Global Positioning System: Theory and Applications
Volume I

Edited by

Bradford W. Parkinson
Stanford University, Stanford, California
James J. Spilker Jr.
Stanford Telecom, Sunnyvale, California

Associated Editors:
Penina Axelrad
University of Colorado, Boulder, Colorado
Per Enge
Stanford University, Stanford, California

Volume 163
PROGRESS IN ASTRONAUTICS AND AERONAUTICS

Paul Zarchan, Editor-in-Chief
Charles Stark Draper Laboratory, Inc.
Cambridge, Massachusetts

Published by the
American Institute of Aeronautics and Astronautics, Inc.
370 L'Enfant Promenade, SW, Washington, DC 20024-2518

Preface

Overview and Purpose of These Volumes

Of all the *military* developments fostered by the recent cold war, the Global Positioning System (GPS) may prove to have the greatest positive impact on everyday life. One can imagine a 21st century world covered by an augmented GPS and laced with mobile digital communications in which aircraft and other vehicles travel through "virtual tunnels," imaginary tracks through space which are continuously optimized for weather, traffic, and other conditions. Robotic vehicles perform all sorts of construction, transportation, mining, and earth moving functions working day and night with no need for rest. Low-cost personal navigators are as commonplace as hand calculators, and every cellular telephone and personnel communicator includes a GPS navigator. These are some of the potential positive impacts of GPS for the future. Our purpose in creating this book is to increase that positive impact. That is, *to accelerate the understanding of the GPS system and encourage new and innovative applications.*

The intended readers and users of the volumes include all those who seek knowledge of GPS techniques, capabilities, and limitations:

- Students attending formal or informal courses
- Practicing GPS engineers
- Applications engineers
- Managers who wish to improve their understanding of the system

Our somewhat immodest hope is that this book will become a standard reference for the understanding of the GPS system.

Each chapter is authored by an individual or group of individuals who are recognized as world-class authorities in their area of GPS. Use of many authors has led to some overlap in the subject matter which we believe is positive. This variety of viewpoints can promote understanding and contributes to our overall purpose. Books written by several authors also must contend with variations in notation. The editors of the volume have developed common notations for the important subjects of GPS theory and analysis, and attempted to extend this, where possible, to other chapters. Where there are minor inconsistencies we ask for your understanding.

Organization of the Volumes

The two volumes are intended to be complementary. Volume I concentrates on fundamentals and Volume II on applications. Volume I is divided into two parts: the first deals with the operation and theory of basic GPS, the second section with GPS performance and errors. In Part I (GPS Fundamentals), a summary of GPS history leads to later chapters which promote an initial understanding of the three GPS segments: User, Satellite, and Control. Even the best of systems has its limitations, and GPS is no exception. Part II, GPS Performance and Error Effects, is introduced with an overview of the errors, followed by chapters devoted to each of the individual error sources.

Volume II concentrates on two aspects: augmentations to GPS and detailed descriptions of applications. It consists of Parts III to VI:

- III. Differential GPS and integrity Monitoring
- IV. Integrated Navigation Systems
- V. GPS Navigation Applications
- VI. Special Applications

Parts III and IV expand on GPS with explanations of supplements and augmentations to the system. The supplements enhance accuracy, availability, or integrity. Of special interest is differential GPS which has proven it can provide sub-meter (even centimeter) level accuracies in a dynamic environment. The last two sections (V and VI) are detailed descriptions of the major applications in current use. In the rapidly expanding world of GPS, new uses are being found all of the time. We sincerely hope that these volumes will accelerate such new discoveries.

Acknowledgments

Obviously this book is a group undertaking with many, many individuals deserving of our sincere thanks. In addition to the individual authors, we would especially like to thank Ms. Lee Gamma, Mr. Sam Pullen, and Ms. Denise Nunes. In addition, we would like to thank Mr. Gaylord Green, Dr. Nick Talbot, Dr. Gary Lennon, Ms. Penny Sorensen, Mr. Konstantin Gromov, Dr. Todd Walter, and Mr. Y. C. Chao.

Special Acknowledgment

We would like to give special acknowledgment to the members of the original GPS Joint Program Office, their supporting contractors and the original set of engineers and scientists at the Aerospace Corporation and at the Naval Research Laboratory. Without their tenacity, energy, and foresight GPS would not be.

B. W. Parkinson
J. J. Spilker Jr.
P. Axelrad
P. Enge

Table of Contents

Part I. GPS Fundamentals

Part II. GPS Performance and Error Effects

Chapter 1

Introduction and Heritage of NAVSTAR, the Global Positioning System

Bradford W. Parkinson*

Stanford University, Stanford, California 94305

I. Background and History

FOR six thousand years, humans have been developing ingenious ways of navigating to remote destination. A fundamental technique developed by both ancient Polynesians and modern navies is the use of *angular measurements* of the natural stars. With the development of radios, another class of navigation aids was born. These included radio beacons, vhf omnidirectional radios (VORs), long-range radio navigation (LORAN), and OMEGA. With yet another technology—artificial satellites—more precise, *line-of-sight* radio navigation signals became possible. This promise was realized in the 1960s, when the U.S. Navy's Navigation Satellite System (known as Transit) opened a new era of navigation technology and capability. However, the best was yet to come.

Over a long Labor Day weekend in 1973, a small group of armed forces officers and civilians, sequestered in the Pentagon, were completing a plan that would truly revolutionize navigation. It was based on radio ranging (eventually with *millimeter precision*) to a constellation of artificial satellites called NAVSTARs. Instead of *angular measurements* to natural stars, greater accuracy was anticipated with *ranging measurements* to the artificial NAVSTARs.

Although it has taken over twenty years to establish that system and to realize its implications fully, it is now apparent that a new *navigation utility* has been created. For under a thousand dollars (price rapidly decreasing), anyone, anywhere in the world, can almost instantaneously determine his or her location in three dimensions to about the width of a street.

This book explains the technology, the performance, and the applications of this new utility—the Global Positioning System (GPS). "*With the quiet revolution of NAVSTAR, it can be seen that these potential uses are limited only by our imaginations*."[1]

*Professor, Department of Aeronautics and Astronautics, and Director, GPS Program.

A. Predecessors

By 1972, the U.S. Air Force and Navy had for several years been competitively studying the possibility of improved navigation from space. These studies became the basis for a new synthesis known as NAVSTAR or the GPS. A brief discussion of the predecessor systems is followed by a description of the Air Force's development program and a summary of the technical design (which is expanded further in succeeding chapters).

1. *Applied Physics Laboratory's Transit: Navy Navigation Satellite System*

The first operational satellite-based navigation system was called NNSS (Navy Navigation Satellite System), or Transit. Developed by the Johns Hopkins Applied Physics Laboratory (APL) under Dr. Richard Kershner, Transit was based on a user measuring the Doppler shift of a tone broadcast at 400 MHz by polar orbiting satellites at altitudes of about 600 nautical miles (actually, two frequencies were transmitted to correct for ionospheric group delay).

The tone broadcast by Transit was continuous. The maximum rate of change in the Doppler shift of the received signal corresponded to the point of closest approach of the Transit satellite. The difference between "up" Doppler and "down" Doppler can be used to calculate the *range to the satellite at closest approach.* Users with known altitude (e.g. sea level) and the broadcast ephemeris of the satellite could use these Doppler measurements to calculate their positions to a few hundred meters. Of course, corrections had to be made for the user's velocity. Because of this velocity sensitivity and the *two*-dimensional nature, Transit was not very useful for air applications. Another limitation was the intermittent availability of the signals, because mutual interference restricted the number of satellites available worldwide to about five. This limited coverage had unavailability periods of 35 to 100 min.

Originally intended as a system to help U.S. submarines navigate, Transit was soon adopted extensively by commercial marine navigators. Although Transit is still operational, new satellites are no longer being launched, and the Federal Radionavigation Plan has announced the intent to phase it out.

Technology developed for Transit has proved to be extremely useful to GPS. Particularly important were the satellite prediction algorithms developed by the Naval Surface Warfare Center. Transit also proved that space systems could demonstrate excellent reliability. After initial "infant mortality" problems, an improved version exhibited operational lifetimes exceeding its specifications by two or three times. In fact, a number of these valuable spacecraft have lasted more than 15 years.

2. *Naval Research Laboratory's Timation Satellites*

By 1972, another Navy satellite system was extending the state of the art by orbiting very precise clocks. Known as *Timation,*[2] these satellites were developed under the direction of Roger Easton at the Naval Research Laboratory (NRL). They were used principally to provide very precise time and time transfer between various points on the Earth. In addition, they could provide navigation information. The ranging signals used a technique called *side-tone ranging,* which broadcast a variety of synchronized tones to resolve phase ambiguities.

Initially these spacecraft used very stable *quartz-crystal oscillators;* later models were to orbit the first *atomic frequency standards* (rubidium and cesium). The atomic clocks typically had a frequency stability of several parts in 10^{12} (per day) or better. This frequency stability greatly improves the prediction of satellite orbits (ephemerides) and also extends the time between required control segment updates to the GPS satellites. Timation satellites were flown in inclined orbits: the first two at altitudes of 500 nautical miles and the last in the series at 7500 nautical miles. The third satellite was also used as a technology demonstrator for GPS (see Fig. 1). This pioneering work in space-qualified time standards was an important foundation for GPS.

3. U.S. Air Force Project 621B

The third essential foundation for GPS was a U.S. Air Force program known as 621B. This program was directed by an office in the Advanced Plans group at the Air Force's Space and Missile Organization (SAMSO) in El Segundo, California. By 1972, this program had already demonstrated the operation of a new type of satellite-ranging signal based on pseudorandom noise (PRN). The signal modulation was essentially a repeated digital sequence of fairly random bits—ones or zeros—that possessed certain useful properties. The sequence could be easily generated by using a shift register or, for shorter codes, simply by storing the entire sequence of bits. A navigation user could detect the start

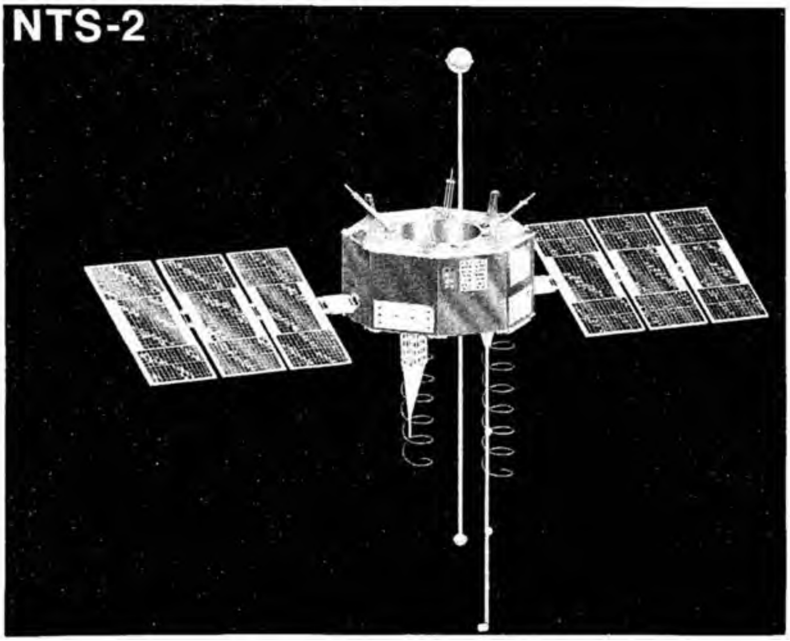


Fig. 1 Navigation technology Satellite II (NTS-II). This satellite was launched as part of the joint program effort to develop reliable spacecraft atomic clocks. NRL called this satellite Timation III (drawing courtesy of NRL).

("phase") of the repeated sequence and use this for determining the range to a satellite. The signals could be detected even when their power density was less than 1/100th that of ambient noise, and *all satellites could broadcast on the same nominal frequency* because properly selected PRN coding sequences were nearly orthogonal. Successful aircraft tests had been run at Holloman AFB to demonstrate the PRN technique. The tests used simulated satellite transmitters located on the floor of the New Mexican desert.

The ability to reject noise also implied a powerful ability to reject most forms of jamming or deliberate interference. In addition, a communication channel could be added by inverting the whole sequence at a slow rate and using these inversions to indicate the ones or zeros of digital data. This slow communication link (50 b/s) allowed the user to receive ephemeris (satellite location) and clock information.

The original Air Force concept visualized several constellations of highly eccentric satellite orbits with 24-h periods. Alternative constellations were nicknamed the *egg-beater, the rotating X and the rotating Y configurations* because of their resulting ground traces. Although these designs allowed the system to be deployed gradually (for example, North and South America first) they had high line-of-sight accelerations. Initially, the concept relied on continuous measurement from the ground to keep the signals time-synchronized. Later, the NRL clock concept was added because the synchronizing link would have been quite vulnerable. The GPS did substitute the Timation clocks later to remove any reliance on continuous ground contact.

B. Joint Program Office Formed, 1973

In the early 1970s, a number of changes in the systems acquisition process had begun to be adopted for the Department of Defense. These changes, recommended by David Packard, were to have a profound effect on NAVSTAR and other major DOD programs. To increase efficiency and reduce interservice bickering, "joint" programs were formed that forced the various services to work together. The GPS was one of the earliest examples. It was decreed to be a Joint Program, with a Joint Program Office (JPO) located at the Air Force's Space and Missile Organization and to have multiservice participation (with the Air Force as the lead service).

The first program director was Dr. (Col.) Bradford W. Parkinson (see Fig. 2), supported by Deputy Program directors—eventually from the Army, Navy, Marine Corps, Defense Mapping Agency, Coast Guard, Air Logistics Command, and NATO. Also continuing their support of 621B was a small cadre of engineers from the Aerospace Corporation under Mr. Walter Melton. Dr. Parkinson was directed to develop the initial concept as a joint development and to gain approval of the Department of Defense to proceed with full-scale demonstration and development.

There have been many speculations on the origin of the names Global Positioning System, and NAVSTAR. The GPS title originated with General Hank Stehling who was the Director of Space for the U.S. Air Force DCS Research and Development (R&D) in the early 1970s. He pointed out to Dr. Parkinson that "navigation" was an inadequate descriptor for the proposed concept. He suggested that "Global

Fig. 2 Joint program included deputy directors from all services. Dr. (Colonel) Parkinson is in discussions with his Navy Deputy, Cdr. Bill Huston of the U.S. Navy. Models of the NTS-II and Phase one GPS satellites are on the table. The civilian is Mr. Frank Butterfield of the Aerospace Corporation (photo courtesy of the U.S. Air Force).

Positioning System" would be a better name. The JPO enjoyed his sponsorship, and this insightful description was immediately adopted.

The title NAVSTAR came into being in a somewhat similar manner. Mr. John Walsh [an Associate Director of Defense Development, Research, and Engineering (DDR&E)] was a key decision maker when it came to the budget for strategic programs in general, including the proposed satellite navigation program. In the contention for funding, his support was not as fervent as the JPO would have liked. During a break in informal discussions between Mr. John Walsh and Col. Brent Brentnall (the program's representative at DOD), Mr. Walsh suggested that NAVSTAR would be a nice sounding name. Colonel Brentnall passed this along as a good idea to Dr. Parkinson, noting that if Mr. Walsh were to name it, he would undoubtedly feel more protective toward it. Dr. Parkinson seized the opportunity, and ever since the program has been known as NAVSTAR the Global Positioning System. Although some have assumed that NAVSTAR was an acronym, in fact, it was simply a nice sounding name* that enjoyed the support of a key DOD decision maker.

*We should note that TRW apparently had advocated a navigational system for which NAVSTAR was an acronym (*NAV*igation *S*ystem *T*iming *A*nd *R*anging). This may have been in Mr. Walsh's subliminal memory, but was not part of the process. It was never used as an acronym.

1. Failed Defense System Acquisition and Review Council

Fortunately, the first attempt to gain system approval failed in August 1973. The program that was brought before the Defense System Acquisition and Review Council (DSARC) at that time was not representative of a joint program. Instead it was packaged as the 621B system. Dr. Malcolm Currie,* then head of DDR&E, expressed strong support for the idea of a new satellite-based navigation system, but requested that the concept be broadened to embrace the views and requirements of all services.

2. Synthesis of a New System

With this philosophy, Dr. Parkinson and the Joint Program Office immediately went to work. Over the Labor Day weekend of 1973, he assembled about a dozen members of the JPO on the fifth floor of the Pentagon. He directed the development of a new design that employed the best of all available satellite navigation system concepts and technology. The result was a system proposal that was not exclusively the concept of any prior system but rather was a synthesis of them all. The details of the proposed GPS are outlined below. Its multiservice heritage precluded any factual basis for further bickering, because all contending parties now were part of the conception process. From that point forward, the JPO acted as a multiservice enterprise, with officers from all services attending reviews and meetings that had previously been "Air Force only."

3. Approval to Proceed with GPS

To gain approval for the new concept, Dr. Parkinson began to contact all those with some stake in the decision. After interminable rounds of briefings† on the new approach were given to offices in the Pentagon and to the operating armed forces, a successful DSARC was held on 17 December 1973. Approval to proceed was granted.

The first phase of the program included four satellites (one was the refurbished qualification model), the launch vehicles, three varieties of user equipment, a satellite control facility, and an extensive test program. By June of 1974, the

*Dr. Currie was appointed head of DDR&E in early 1973, as part of the incoming administration. He had been living in Los Angeles prior to his appointment and to complete his move he made numerous trips to Los Angeles in the initial months. One legitimate official purpose of these trips was to review programs at SAMSO. After a few trips, he had done all the high-level reviews that were available, so the head of SAMSO, General Schultz, suggested that he receive an indepth review by Dr. (Col.) Parkinson on the space-based navigation concept, then known as 621B. This resulted in a remarkable meeting with the number-three man in all of the U.S. DOD spending about three hours in a small office with a lowly Colonel, talking about engineering, technology, and the wide applications of the proposed system. With his doctorate in Physics, Mal Currie was a keen and quick study. He had a great deal of space experience from his years at Hughes Aircraft. The outcome was that the GPS program enjoyed his steadfast support. Without this key decision maker, the Air Force would have killed the program in favor of additional airplanes. The pivotal (and coincidental) meeting with Dr. Parkinson was destined to be an essential factor in gaining system approval.

†Lt. General Ken Schultz was particularly incensed with the endless presentations that had to be made in the Washington arena. The situation with any bureaucracy is that many can say no, and few (if any) can say yes. To bring the naysayers to neutral, the extended trips from Los Angeles to Washington were necessary for Dr. Parkinson.

satellite contractor, Rockwell International, had been selected, and the program was well underway. The initial types of user equipment included sequential and parallel military receivers, as well as a civil type set for utility use by the military. The development test and evaluation was extensive, with a laser-tracking range set up at the Army's Yuma Proving Ground. An independent evaluation was performed by the Air Force's Test and Evaluation Command.

To maintain the focus of the program the JPO adopted a motto:

> The mission of this Program is to:
> 1. Drop 5 bombs in the same hole, and
> 2. Build a cheap set that navigates (<$10,000),
>
> and don't you forget it!

The program developed rapidly; the first operational prototype satellite was launched in February of 1978 (44 months after contract start). By this time, the initial control segment was deployed and working, and five types of user equipment were undergoing preliminary testing at the Yuma Proving Ground. The initial user equipment types had been expanded to include a 5-channel set developed by Texas Instruments and a highly jam-resistant set developed by Rockwell Collins.

4. *Needed: A Few More Good Satellites*

As stated, only four satellites were initially approved by the DOD, including a refurbished qualification model (see Fig. 3). It became apparent that there was a need for additional satellites, because the minimum number for three-dimensional navigation is four. Any launch or operational failure would have gravely impacted the Phase I demonstration program. Authorization for spare GPS satellites was urgently needed.

The Navy's Transit program inadvertently solved this problem. The chain of events began when Transit requested funds for upgrading certain Transit satellites to a PRN code similar to that used by GPS. The purpose was to provide accurate tracking of the Trident (submarine launched missile) booster during test firings into the broad ocean areas. Dr. Bob Cooper of DDR&E requested a series of reviews to address whether GPS could fulfill this mission.

The GPS solution was to use a signal translator on the Trident missile bus that would relay the GPS modulations to the ground on another frequency. The central issues were whether the ionosphere could be adequately calibrated (because it was a single-frequency system, the ionosphere could not be directly measured), and whether the translated signal could be recorded with sufficient fidelity (it required digitizing at 60 MHz).

During the third and capstone review for Dr. Cooper, Dr. Parkinson (supported by Dr. Jim Spilker and Dr. Jack Klobuchar) was able to present convincing arguments that a GPS solution could solve the Trident problem *provided two additional satellites were authorized.* Dr. Cooper immediately made the decision to use GPS. He directed the transfer of $60M from the Navy to the Air Force, approving two additional satellites, and thereby greatly expanding the Phase one test time as well as significantly reducing the program risk. This little known

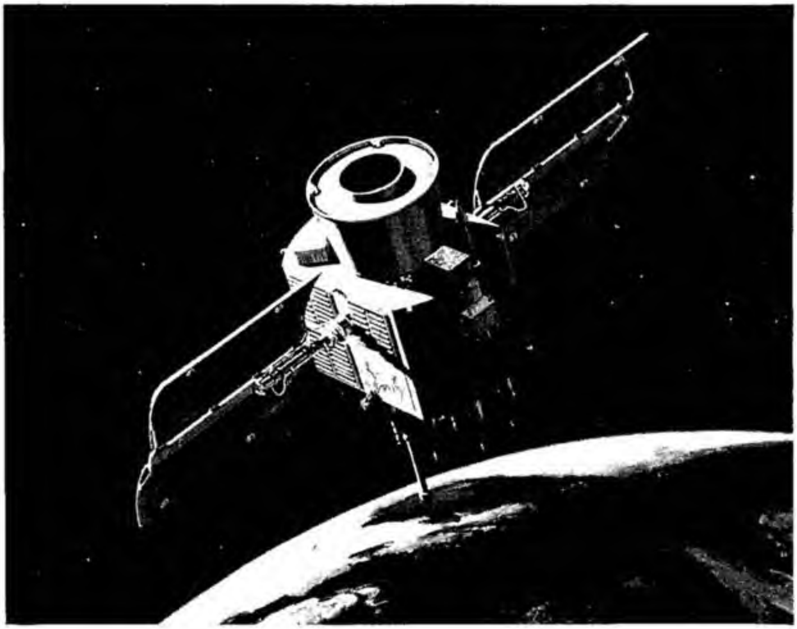


Fig. 3 Phase I GPS satellite. It is a three-axis stabilized design with double and triple redundancy where appropriate (drawing courtesy of the U.S. Air Force).

event also eliminated the possibility of an upgraded Transit program competing with the fledgling GPS.

III. Introductory GPS System Description and Technical Design

The operational GPS system of today is virtually identical to the one proposed in 1973. The satellites have expanded their functionality to support additional military capabilities; the orbits are slightly modified, but the equipment designed to work with the original four satellites would still perform that function today. The rest of this volume is devoted to detailed technical descriptions of the system and its applications; the following section provides an overview of the system design.

A. Principles of System Operation

The fundamental navigation technique for GPS is to use *one-way ranging* from the GPS satellites that are also broadcasting their estimated positions. *Ranges are measured to four satellites simultaneously in view by matching (correlating) the incoming signal with a user-generated replica signal and measuring the received phase against the user's (relatively crude) crystal clock.* With four satellites and appropriate geometry, four unknowns can be determined; typically, they are: latitude, longitude, altitude, and a correction to the user's clock. If altitude or time are already known, a lesser number of satellites can be used (see Fig. 4).

Each satellite's future position is estimated from ranging measurements taken at worldwide monitoring stations.* These ranging measurements use the same signals that are employed by a typical user's receiver. Using sophisticated prediction algorithms, the master control station forms estimates of future satellite locations and future satellite clock corrections. For the uploads, which occur daily or (optionally) more frequently, the combined predictions for satellite clock and position have been measured to have an average rms error of 2–3 m. These estimates have demonstrated reasonable errors even after three days (24.3 m of expected ranging error).

B. GPS Ranging Signal

The GPS ranging signal is broadcast at two frequencies: a primary signal at 1575.42 MHz (L_1) and a secondary broadcast at 1227.6 MHz (L_2). These signals are generated synchronously, so that a user who receives both signals can directly calibrate the ionospheric group delay and apply appropriate corrections. However, most civilian users will only use the primary or L_1 frequency.

Potentially, both the signal at the L_1 frequency and the signal at L_2 can *each* have two modulations at the same time (called "phase quadrature"). Current

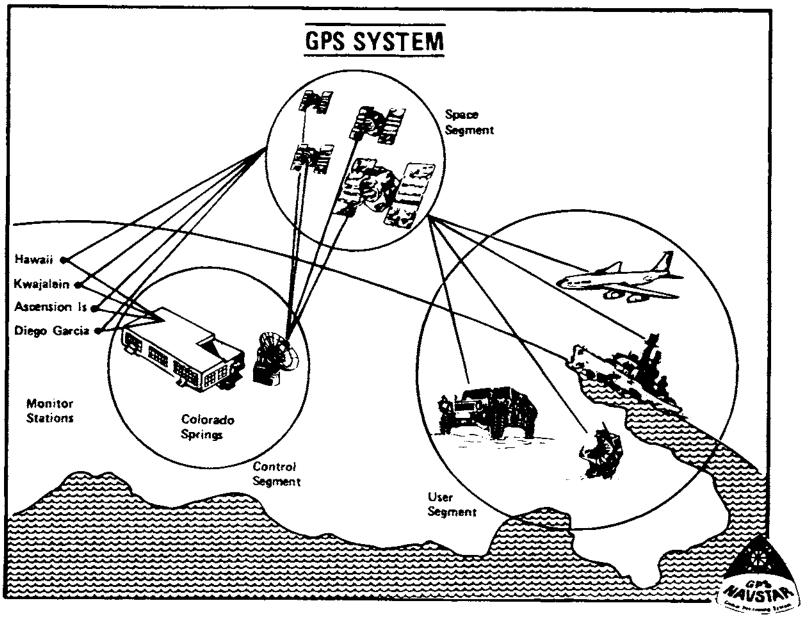


Fig. 4 System configuration of GPS showing the three fundamental segments: 1) user; 2) spacecraft; and 3) ground control (drawing courtesy of the U.S. Air Force).

*The Operational Control System (OCS) uses five monitor stations which are located at Colorado Springs, Ascension Island, Diego Garcia, Kwajalein, and Hawaii.

implementation has two modulations on the higher frequency (L_1), but only a single (protected) modulation (see below) on L_2 The two modulations are as follows:

1) C/A or Clear Acquisition Code: This is a short PRN code broadcast at a bit (or chipping) rate of 1.023 MHz. This is the principal civilian ranging signal, and it is always broadcast in the *clear* (unencrypted). It is also used to acquire the much longer P-code. The use of this signal is called the Standard Positioning Service or SPS. It is always available, although it may be somewhat degraded. At this time, and for the projected future, the C/A code is available only on L_1 (some civil users have requested C/A modulation on L_2 to allow ionospheric calibration).

2) P or Precise Code (sometimes called the Protected Code): A very long code (actually segments of a 200-day code) that is broadcast at ten times the rate of C/A, 10.23 MHz. Because of its higher modulation bandwidth, the code ranging signal is somewhat more precise. This reduces the noise in the received signal but will not improve the inaccuracies caused by biases. This signal provides the Precise Positioning Service or PPS. The military has encrypted this signal in such a way that renders it unavailable to the unauthorized user. This ensures that the unpredictable code (to the unauthorized user) cannot be spoofed. This feature is known as antispoof or AS. When encrypted, the P code becomes the Y code. Receivers that can decrypt the Y code are frequently called P/Y code receivers. As a result of the military intent, most civilian users should only rely on the C/A code or SPS.*

1. *Selective Availability*

In addition, the military operators of the system have the capability to degrade the accuracy of the C/A code intentionally by desynchronizing the satellite clock, or by incorporating small errors in the broadcast ephemeris. This degradation is called *Selective Availability, or S/A.* The magnitude of these *ranging* errors is typically 20 m, and results in rms *horizontal position* errors of about 50 m, one sigma. The official DOD position is that errors will be limited to 100 m, 2 drms, which is about the 97th percentile. A technique known as differential GPS (explained later) can overcome this limitation and potentially provide accuracies sufficient for precision approach of aircraft to landing fields.

2. *Data Modulation*

One additional feature of the ranging signal is a 50 b/s modulation used as a communications link. Through this link, each satellite transmits its location and the correction necessary to apply to the spaceborne clock.† Also communicated are the health of the satellite, the locations of other satellites, and the necessary information to lock on to the P code after acquiring the C/A code.

*There are provisions in the Federal Radionavigation Plan for civilian users with critical national needs to gain access to the P code.

†Although the atomic clocks are extremely stable, they are running in an uncorrected mode. The clock correction is an adjustment that synchronizes all clocks to GPS time.

C. Satellite Orbital Configuration

The orbital configuration approved at DSARC in 1973 was a total of 24 satellites—eight in each of three circular rings with inclinations of 63 deg. The rings were equally spaced around the equator, and the orbital altitudes were 10,980 n.mi. This altitude gave two orbital periods per sidereal day (known as *semisynchronous*) and produced repeating ground traces. The altitude was a compromise among: user visibility, the need to pass over the continental U.S. ground/upload stations periodically, and the cost of the spacecraft launch boosters. Three rings of satellites were initially selected because it would be easier to have orbital spares—only three such spares could easily replace any single failure in the whole constellation. This configuration provided a minimum of six satellites in view at any time, with a maximum of 11. As a result of this redundancy, the system was robust in the sense that it could tolerate occasional satellite outages (see Fig. 5).

Two changes have been made since the original constellation proposal. The inclinations have been reduced to 55 deg, and the number of orbital planes have been increased to six,* with four satellites in each. The number of satellites, including spares, remains 24.

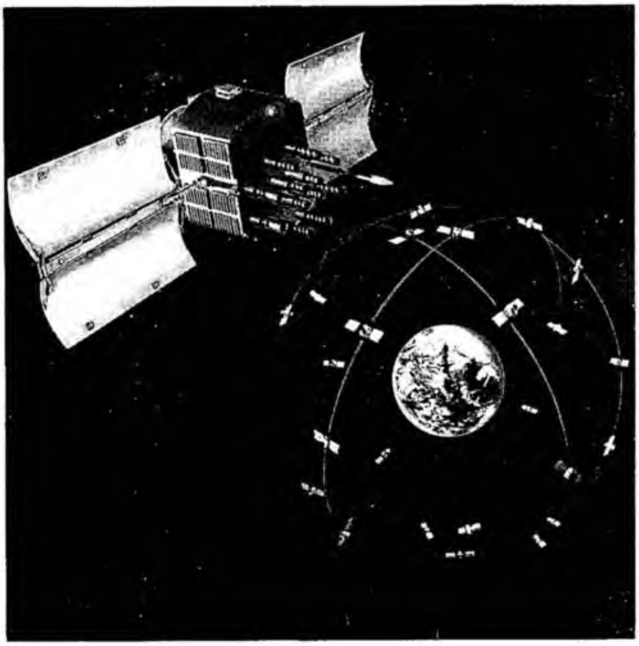


Fig. 5 Original GPS orbital configuration of three rings of eight satellites each. The final operational configuration has the same number of satellites, arranged in six rings of four satellites (drawing courtesy of the U.S. Air Force).

*For a number of reasons, the Department of Defense calls the configuration "21 satellites with three orbiting spares." We may find the system eventually having more satellites to increase robustness.

D. Satellite Design

The GPS satellites are attitude stabilized on all three axes and use solar panels for basic power (see Fig. 6). The ranging signal is radiated through a shaped beam antenna—by enhancing the received power at the limbs of the Earth, compensation is made for "space loss." The user, therefore, receives fairly constant power for all local elevation angles.* The satellite design is generally doubly or triply redundant, and the Phase I satellites demonstrated average lifetimes in excess of 5 years (and in some cases over 12).

E. Satellite Autonomy: Atomic Clocks

A key feature of the GPS design is that the satellites need not be continuously monitored and controlled. To achieve this autonomy, the satellites must be predictable in four dimensions: three of *position* and one of *time.* Predictability, in the orbital *position,* is aided because the high-altitude orbits are virtually unaffected by atmospheric drag. Many other factors which affect orbital position must also be considered. For example, variations in geopotential, solar pressure, and outgassing can all have significant effects.

When GPS was conceived, it was recognized that the most difficult technology problem facing the developers was probably the need to fly accurate timing

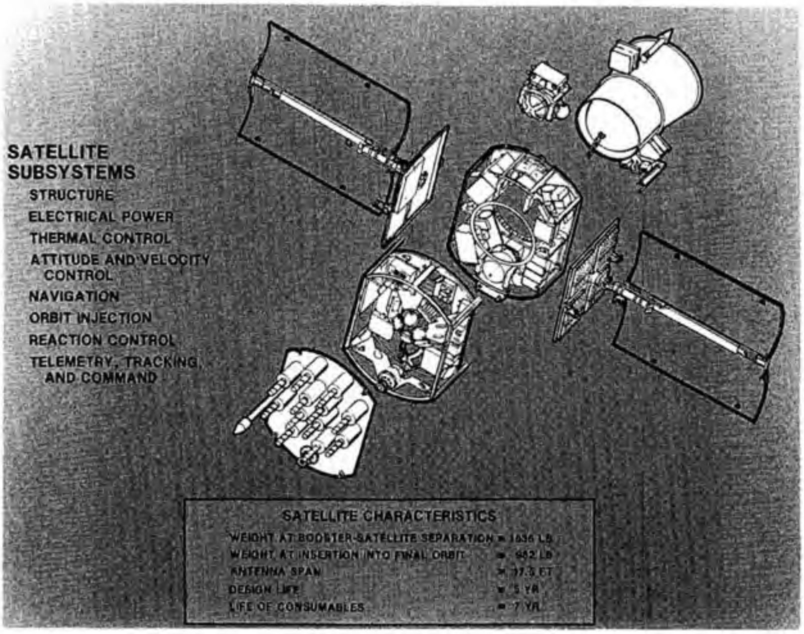

SATELLITE SUBSYSTEMS
STRUCTURE
ELECTRICAL POWER
THERMAL CONTROL
ATTITUDE AND VELOCITY CONTROL
NAVIGATION
ORBIT INJECTION
REACTION CONTROL
TELEMETRY, TRACKING, AND COMMAND
SATELLITE CHARACTERISTICS
WEIGHT AT BOOSTER-SATELLITE SEPARATION = 1636 LB
WEIGHT AT INSERTION INTO FINAL ORBIT = 962 LB
ANTENNA SPAN = 17.5 FT
DESIGN LIFE = 5 YR
LIFE OF CONSUMABLES = 7 YR

Fig. 6 Breakaway view of the GPS Phase I satellite design (drawing courtesy of the U.S. Air Force).

*The requirement for received power on L_1 is -163 dbw into an isotropic, circularly polarized antenna on the primary frequency.

standards, insuring that all satellites' clocks remained synchronized. As mentioned, NRL had been developing frequency standards for space, so this effort was continued and extended.

Payoff of a Good Clock

The basic arithmetic can be understood as follows: A day is about 100,000 s, or 10^5. Light travels about 1 ft per ns (10^{-9} s). If the system can tolerate an error buildup caused by the atomic clock of 5 ft, the stability must be 5 ns per upload (one-half a day). This is about $(5*10^9)/(5*10^4)$ sps, measured over 12 h. Therefore, this requirement is for a clock with about one part in 10^{13} stability,* which can only be met by an atomic standard. Note that there is a roughly constant frequency shift attributable to relativistic effects (both special and general) of about 4.5 parts in 10^{10}, which is compensated by a deliberate offset in the clock frequency.

GPS traditionally has used two types of atomic clocks: rubidium and cesium. Phase one test results for the rubidium cell standard are shown in Fig. 7. A key to outstanding satellite performance has been the stability of the space-qualified atomic clocks, which exceeded the specifications. They have measured stabilities of one part in 10^{13} over periods of 1–10 days.[3]

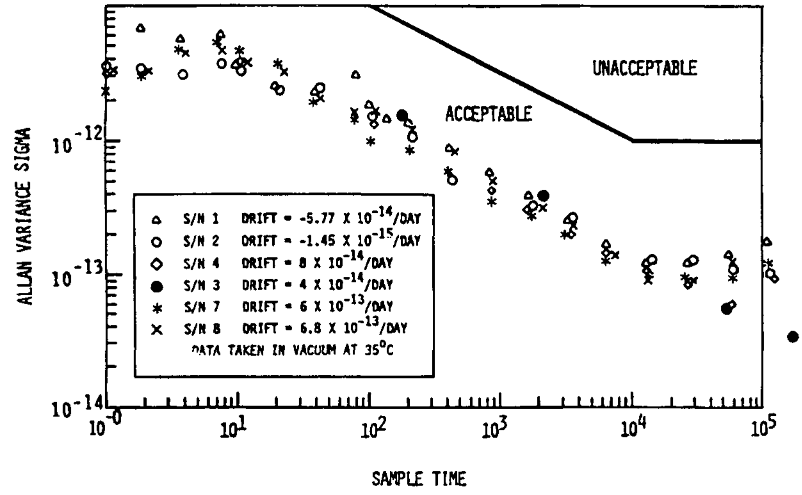


Fig. 7 Space qualified rubidium-cell frequency standard performance. These units were developed by Rockwell as a derivative of a clock designed by Efratom, Inc. (data courtesy of the U.S. Air Force).

*Clock stability is traditionally measured with the Allen variance, which shows stability versus averaging time. For short averaging times (1 s) virtually all clocks are dominated by the quartz oscillator, which acts as the short-term flywheel. In Phase one, the clocks were specified at 10^{-12}, measured over 1 day.

F. Ionospheric Errors and Corrections

The free electrons in the ionosphere create a delay in the modulation signal (PRN code). This delay is proportional to the integrated number of free electrons along the transmission path and inversely proportional to the square of the transmission frequency (to first order). The path delay at any elevation angle is often expressed as the product of a zenith delay (elevation equals 90 deg) and an *obliquity* factor that is a function of that elevation angle. This ratio ranges from 1.0 at the zenith (by definition) to about 3.0 at small elevation angles. Typical zenith (or vertical) delays range from a few meters at night to a maximum of ten or twenty meters at about 1400 hours (local solar time). Thus, it is not unusual to find delays of over 30 m at lower elevation angles. Fortunately, these delays are highly correlated between satellites, which helps reduce the calculated horizontal position errors.*

There are two techniques for correcting this error. The first is to use an ionosphere model. The 8-model parameters used to calculate the correction are broadcast as part of the GPS 50 b/s message. This model is typically accurate to a few meters of vertical error.

The second technique uses both broadcast frequencies and the inverse square law behavior to measure the delay directly. By differencing the code measurements on each frequency, the delay on L_1 is approximately 1.546*(difference in delays on L_1 and L_2). This technique is only available to a P/Y-code receiver (because the only L_2 modulation is the P code) or to a codeless (or cross-correlating) receiver.

G. Expected Navigation Performance

The performance capabilities of GPS are primarily affected by two things: 1) the *satellite geometry* (which causes *geometric dilution*); and 2) the *ranging errors*. Under the assumption of uniform, uncorrelated, zero-mean, ranging-error statistics, this can be expressed as follows:

$$\boxed{\text{RMS position error} = (\text{Geometric dilution}) * (\text{rms ranging error})}$$

1. Geometric Dilution

The geometric dilution can be calculated for any instantaneous satellite configuration, as seen from a particular user's location. The details of this calculation are explained in Chapters 6 and 11, this volume. For a 21-satellite constellation and a three-dimensional fix, the world median value of the geometric dilution factor (for the nominal constellation) is about 2.7. This quantity is usually called PDOP or position dilution of precision. Typical dilution factors range from 1.5 to 8. The *variations* in this *dilution factor* are typically much greater than the variations in ranging errors.

*This correlation of errors due to the ionosphere will mostly show up as a user clock error. While this is not important for many navigation users, it is critical for the precise transfer of time. Such users must employ the dual frequency technique to eliminate this error.

2. *Ranging Errors*

Ranging errors are generally grouped into the following six major causes:

1) Satellite ephemeris
2) Satellite clock
3) Ionospheric group delay
4) Tropospheric group delay
5) Multipath
6) Receiver measurement errors, including software

Some of these errors tend to be correlated for the same satellite. For example, satellite clock and ephemeris errors tend to be negatively correlated; i.e. they tend to cancel each other somewhat. Other errors tend to be correlated between satellites. For example, the ionospheric and tropospheric group delays always have the same sign, because they are the result of signal paths penetrating the *same blanket of media* with different angles.

With S/A turned off, all errors for *single frequency* SPS are nearly *identical* in magnitude to those for *single-frequency PPS* except for receiver measurement errors (which decrease with increasing bandwidth). *Dual* frequency, which is only available on PPS, can reduce the third error (attributable to the ionosphere) to about 1 m.*

3. *Precise Positioning Service (PPS) Accuracy*

Ranging errors (including the effects of the satellite clock)[3] for the PPS have been *specified* to be better than 6 m. The product of the average PDOP and the ranging error is the specified three-dimensional accuracy of 16 m spherical error probable (SEP).†

Because each of the five worldwide monitoring stations is continuously measuring the ranging errors to all satellites in view, these measurements are a convenient statistic of the basic, static accuracy of GPS. Table 1 summarizes over 11,000 measurements taken from 15 January to 3 March 1991, during the "Desert Storm" operation of the Gulf War. The S/A feature was not activated during this period. Note that the PPS results presumably are not affected by S/A at all.

During this period, satellite (PRN 9) was ailing but is included in the solution, making the results somewhat worse than would be expected. By dividing the overall SEP by the rms PDOP, an estimate of the effective ranging error can be formed. The average of these results is 2.3 m.‡ This should be compared to the specification of 6 m. Because SEP is smaller than the RMS error, this estimate may be about 15% optimistic.

*Multipath errors are generally negligible for path delays that exceed one-and-one-half modulation chips, expressed as a range. Thus, P-code receivers reject reflected signals whose path delay exceeds 150 feet. For the C/A code, the number is 1500 feet, giving a slight advantage to the P code, although it is usually reflections from very close objects that are the main source of difficulty.

†SEP is the radius of the sphere that will contain 50% of the expected errors in three dimensions.

‡This number is probably somewhat better than an average receiver would measure for several reasons. Monitor station receivers are carefully sited to avoid multipath. The receivers are of excellent quality and are not moving. Also, since the monitor station measurements are used to update the ephemeris, there may be some tuning to make the predictions match any peculiarities (e.g. survey errors) at the monitor station locations. Nonetheless, an average ranging error of 2.3 m is an impressive result.

Table 1 PPS measured accuracies; SEP/CEP[a] navigation errors; PPS solutions for the OCS monitor stations during Desert Storm; S/A is off.[4]

Criteria	All	Colorado Springs	Ascension	Hawaii	Diego Garcia	Kwajalein
SEP three-dimensional, m	8.3	7.8	6.8	9.0	9.1	9.0
CEP two-dimensional, m	4.5	4.5	3.8	5.1	4.6	5.0
rms PDOP	3.6	3.9	3.4	3.9	3.4	3.3
Estimated range error, m[b]	2.3	2.0	2.0	2.3	2.7	2.7

[a] CEP is circular error probable, which equals the radius of a circle that would contain 50% of the errors. It is the two-dimensional analog of SEP.

[b] This row is formed by dividing the SEP by the rms PDOP.

Table 2 Expected accuracies for various operating conditions of GPS

	Precise positioning service (PPS)		Standard positioning, service (SPS), estimated capability	
	Specification	Measured, static	No S/A	With S/A
Ranging accuracy	6 m	2.3 m	6 m	20 m
CEP (horizontal)	——	4.6 m	12 m	40 m
SEP (three-dimensional)	——	8.3 m[a]	22 m	72 m

[a] The SEP reported for dynamic PPS users was less than 10 m. See Ref. 5.

a. SPS accuracy. Without the degradation of S/A, the SPS would provide solutions with about 50% greater error because of uncompensated ionospheric effects and somewhat greater receiver noise (because of the narrower band C/A code). *It is reasonable to expect that rms horizontal errors for SPS with S/A off would be less than 10 m.*

b. Accuracy summary for code-tracking receivers. Table 2 summarizes the expected accuracies for GPS.

H. High Accuracy/Carrier Tracking

A special feature of GPS, which initially was not generally understood, is the ability to create an extremely precise ranging signal by reproducing and tracking the rf carrier (1575.42 MHz). Because this signal has a wavelength of 19 centimeters (7.5 inches), tracking it to 1/100th of a wavelength provides a precision of about 2 mm. Modern receivers can attain these tracking precisions, but unfortunately, this is not *accuracy.* To provide equivalent accuracy, we must determine which carrier cycle is being tracked (relative to the start of modulation) and compare this with another carrier tracking receiver located at a known position.

Surveyors use a technique of double or triple differencing to resolve this *cycle ambiguity*. For dynamic users, the problem is a bit harder. Reflected signals (multipath) and distortions of the ionosphere can be significant errors.

Generally, the carrier-tracking techniques can be used in two ways. For normal use, the carrier tracking can smooth the code tracking and greatly reduce the noise content of the ranging measurement. The other use is in a differential mode for which there are several variations, including surveying, direct measurement of vehicle attitude (with multiple antennas), and various forms of dynamic differential. The GPS control segment uses accumulated delta range (ADR) as the measurement for the monitor stations. This is done with incremental counts of carrier cycles. Later chapters in this volume discuss these techniques in greater detail.

III. History of Satellites

Five groups (or blocks) of satellites have been developed for the GPS program. In chronological order they are: 1) Navigation technology satellites (NTS); 2) navigation development satellites—Block I or NDS; 3) block II satellites; 4) block IIA satellites; and 5) block IIR satellites. In addition, a follow-on group, called block IIF, is being planned.

A. Navigation Technology Satellites

The satellites of the first group were used to explore space technology. They were an extension of the Timation program of the NRL and were known as NTS, or, Navigation Technology Satellites. The first, NTS-1, had been planned as Timation II and was renamed when the JPO was formed. It was launched into a lower orbit than GPS (7500 n.mi.) on the 14th of July 1974. It was the first to fly atomic clocks: two rubidium oscillators were included. The second (and last) of the series, NTS-2, orbited a number of payload components that were identical to the development GPS satellites. This satellite included the first cesium clock in space, the PRN code generator used in the next block, and the first GPS spaceborne computer. The last two items were developed by Rockwell International under the JPO.

B. Navigation Development Satellites—Block I

These pioneering satellites were developed by Rockwell International for the JPO. The initial buy was for four, followed by two additional to support the Trident program. Later, six more were purchased as replacements. Of the 11 satellites that made it into orbit, all achieved initial operational capability. The sole premature failure in these satellites [in all, the satellites launched through the initial Operational capability (IOC)] was caused by the malfunction of a refurbished Atlas-F booster at Vandenberg. The first launch was on the 22nd of February 1978, 44 months after contract award. Designed for 3-year lifetimes, several operated for over 10 years.

C. Operational Satellites—Block II and IIA

In all, 29 satellites have been produced and are on orbit or ready to launch. The first satellite in this series was declared operational on the 10th of August

1989. These satellites were initially launched at a rate of about six per year. Initial operational capability was declared at the end of 1993, with full operational capability (FOC) attained by the end of 1994.

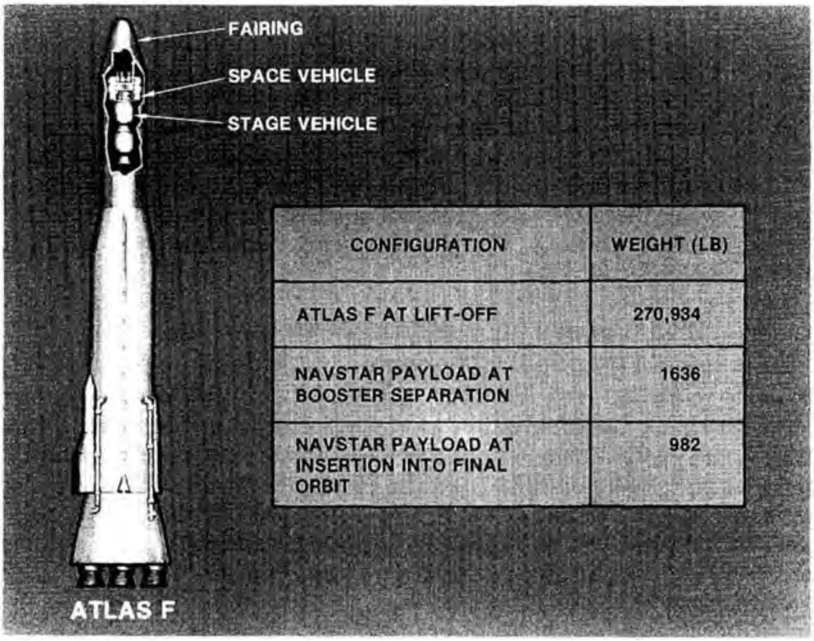
D. Replacement Operational Satellites—Block IIR

These are enhanced performance GPS satellites being developed by Martin Marietta (after buying out the division from GE, which bought it from RCA). The contract was awarded on 21 June 1989 for 20 satellites, with options for 6 more. The first delivery is planned for 1995. These satellites have enhanced autonomy, including the capability to meet a degraded range error specification of up to 180 days since the last ground control segment upload. They also have increased hardening against natural and man-made radiation.

IV. Launches

A. Launch Vehicles

The 12 original (Phase I) satellites were to be launched on refurbished Atlas-F ICBMs (see Fig. 8). The initial plan was to use the McDonnell-Douglas Delta for the next series of launches (Phase II). About 1979, this was changed, and the shuttle was decreed to be the booster of choice for Air Force missions. The



CONFIGURATION	WEIGHT (LB)
ATLAS F AT LIFT-OFF	270,934
NAVSTAR PAYLOAD AT BOOSTER SEPARATION	1636
NAVSTAR PAYLOAD AT INSERTION INTO FINAL ORBIT	982

Fig. 8 Refurbished Atlas F launch vehicle used for the first 12 launches. The stage vehicle was a tandem (stacked) solid rocket configuration (drawing courtesy of the U.S. Air Force).

Table 3 History of initial GPS launches through IOC

Block II seq.	SVN	PRN, code	Internat. ID	NASA catalog number[a]	Orbit plane pos'n.[b]	Launch date, UTC	Clock[c]	Available	Decommissioned
Block I									
——	01	04	1978-020A	10684	——	78-02-22	——	78-03-29	85-07-17
——	02	07	1978-047A	10893	——	78-05-13	——	78-07-14	81-07-16
——	03	06	1978-093A	11054	——	78-10-06	——	78-11-13	92-05-18
——	04	08	1978-112A	11141	——	78-12-10	——	79-01-08	89-10-14
——	05	05	1980-011A	11690	——	80-02-09	——	80-02-27	83-11-28
——	06	09	1980-032A	11783	——	80-04-26	——	80-05-16	91-03-06
——	07		——	——	——	81-12-18	Launch failure		——
——	08	11	1983-013A	14189	——	83-07-14	——	83-08-10	93-05-04
——	09	13	1984-059A	15039	C-1	84-06-13	Cs	84-07-19[d]	——
——	10	12	1984-097A	15271	A-1	84-09-08	Rb	84-10-03	——
——	11	03	1985-093A	16129	C-4	85-10-09	Rb[e]	85-10-30	——
——	——	——	——	——	——	——	——	——	——
Block II									
II-1	14	14	1989-013A	19802	E-1	89-02-14	Cs	89-04-15	——
II-2	13	02	1989-044A	20061	B-3	89-06-10	Cs	89-08-10	——
II-3	16	16	1989-064A	20185	E-3	89-08-18	Cs	89-10-14	——
II-4	19	19	1989-085A	20302	A-4	89-10-21	Cs	89-11-23	——
II-5	17	17	1989-097A	20361	D-3	89-12-11	Cs	90-01-06	——
II-6	18	18	1990-008A	20452	F-3	90-01-24	Cs	90-02-14	——
II-7	20	20	1990-025A	20533	B-2	90-03-26	Cs	90-04-18	——
II-8	21	21	1990-068A	20724	E-2	90-08-02	Cs	90-08-22	——
II-9	15	15	1990-088A	20830	D-2	90-10-01	Cs	90-10-15	——
——	——	——	——	——	——	——	——	——	——
Block IIA									
II-10	23	23	1990-103A	20959	E-4	90-11-26	Cs	90-12-10	——
II-11	24	24	1991-047A	21552	D-1	91-07-04	Cs	91-08-30	——
II-12	25	25	1992-009A	21890	A-2	92-02-23	Rb	92-03-24	——
II-13	28	28	1992-019A	21930	C-2	92-04-10	Cs	92-04-25	——
II-14	26	26	1992-039A	22014	F-2	92-07-07	Cs	92-07-23	——
II-15	27	27	1992-058A	22108	A-3	92-09-09	Cs	92-09-30	——
II-16	32	01[f]	1992-079A	22231	F-1	92-11-22	Cs	92-12-11	——
II-17	29	29	1992-089A	22275	F-4	92-12-18	Cs	93-01-05	——
II-18	22	22	1993-007A	22446	B-1	93-02-03	Cs	93-04-04	——
II-19	31	31	1993-017A	22581	C-3	93-03-30	Cs	93-04-13[g]	——
II-20	37	07	1993-032A	22657	C-4	93-05-13	Cs	93-06-12	——
II-21	39	09	1993-042A	22700	A-1	93-06-26	Cs	93-07-20	——
II-22	35	05	1993-054A	22779	B-4	93-08-30	Cs	93-09-28	——
II-23	34	04	1993-068A	22877	D-4	93-10-26	Cs	93-11-22	——
II-24	——	——	——	——	——	——	——	——	——

[a] NASA Catalog Number is also know as NORAD or U.S. Space Command object number.

[b] No orbital plane position listed = satellite no longer operational.

[c] Clock: Rb = rubidium; Cs = cesium

[d] The power supply of PRN 13 has insufficient capacity to maintain L_1/L_2 transmissions during eclipse season. During this period, the L_1/L_2 transmissions of PRN 13 may be turned off for up to 12 h a day.

[e] PRN 03 is operating on Rb clock without temperature control.

[f] The PRN number of SVN 32 was changed from 32 to 01 on 93-01-28.

[g] Corrective maintenance performed on PRN 31 on 93-06-16 seems to have fixed the L_2 intermittent-lock problem.

Block II satellites were designed to that interface. After the Challenger accident, this decision was reconsidered, and the Delta II has since been used as the GPS launch vehicle. The history and recent status of launches is shown Table 3.

V. Initial Testing

The objective of the Phase I approval of GPS was to validate the total system concept. A major stumbling block in obtaining Phase I approval was a classic bureaucratic "catch 22." The issues raised were the following: 1) How could user equipment development be approved when it wasn't clear they would work with the satellites? but . . . 2) How could the satellites be launched without ensuring they would work with the user equipment? Pursued to a superficial conclusion, nothing could be done at all. The solution was adapted from the 621B program. A system of solar-powered GPS transmitters was deployed on the desert floor at the Yuma test ground. These transmitters all radiated one of the unique orthogonal GPS codes (at the approved frequencies), which were synchronized to each other and to the satellites as they were launched. These transmitters were called pseudolites (from pseudosatellites). They provided a geometry that approximated that of the satellites, although the signals were coming from negative elevation angles. The user equipment could be verified to work with satellite transmitters prior to launch.* As satellites were launched, psuedolites could be dropped from the test system; when four satellites were on orbit, the equipment was completely debugged and able to verify the claims that had been made at DSARC. This approach solved the logical impasse. The pseudolite concept was later expanded as a technique to improve accuracy and integrity for civil landing of aircraft.

This approach satisfied the doubters, and in fact significantly strengthened the program. By 1978, when the first NDS satellite was launched, the main varieties of user equipment had been validated quantitatively and qualitatively.

A. Test Results

Initial testing of user equipment included seven different types that were integrated into 11 types of land, sea, and air vehicles plus manpack testing. Literally hundreds of tests were run. Two results are presented here. Figure 9 shows the summary of integrated tests. Values of SEP† range from 6 to 16 m. The later testing in the A-6 is probably more representative of total system performance.

The blind bombing results, Fig. 10, reflect a substantial improvement because of the GPS. Radar bombing is the usual alternative in poor visibility. The results are particularly significant because the probability of a hit is usually inversely

*In fact a further cross check had been conceived and implemented by (then Major) Gaylord Green who initially ran the satellite development for the JPO. The satellite transmitter was activated during ground testing and shown to allow lock up by the phase one user equipment. Col. Green later returned to the JPO and completed a distinguished career as the Director of the GPS program.

†Spherical error probable is the radius of the sphere that contains 50% of the measured errors. The A-6, B-52, and F-16 are aircraft. The SSN is a submarine, the CV is an Aircraft Carrier. The UH-60 is a helicopter, and MV is the manpack/vehicular set test. Data for these two sets of results come from the JPO publication YEE-82-009B of September 1986, titled "User's Overview" (unclassified).

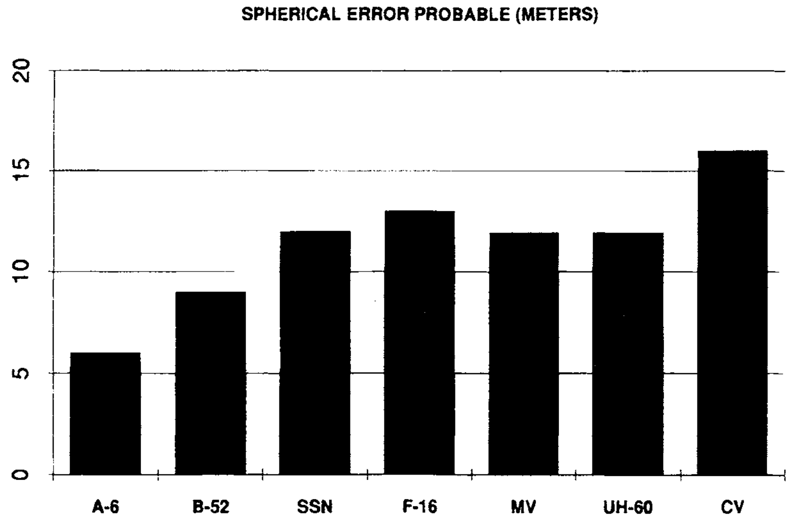


Fig. 9 Test results for seven vehicles using integrated GPS. The earliest tests are on the right and the later on the left. The improved accuracy, in part, reflects system maturity (data courtesy of the U.S. Air Force).

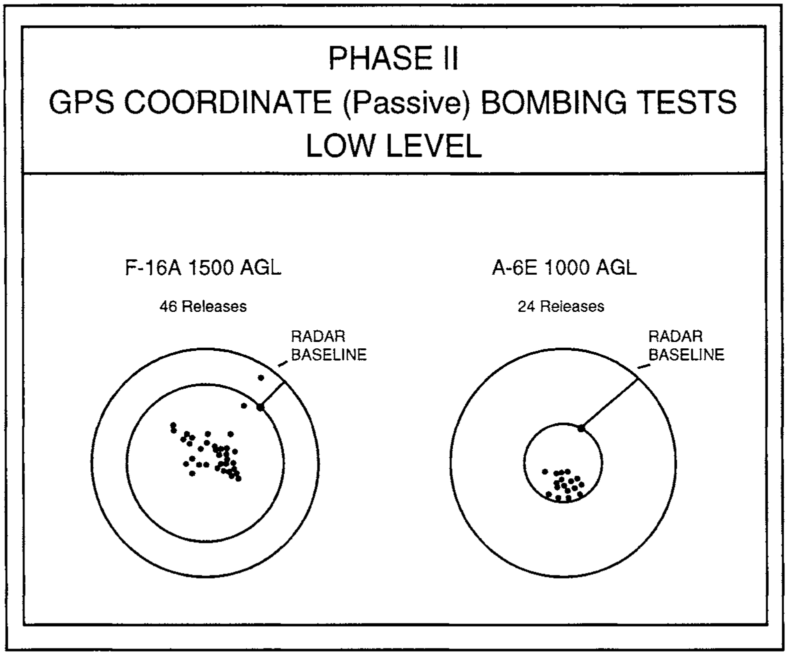


Fig. 10 GPS blind bombing results compared to radar bombing baselines (data courtesy of the U.S. Air Force).

proportional to the *square of the miss radius*. Indeed these results more than satisfied the original part of the motto that called for "five bombs in the same hole."

B. Conclusions

During its initial tests, the Global Positioning System more than met the original design objectives. The impact was not fully appreciated by the operational forces of the military until the Desert Storm battles showed the value of GPS as a force multiplier. The second half of the motto "build a cheap set ($10,000 1973 dollars) that navigates has been overtaken by the civil rush to build integrated chip sets that have driven the costs of GPS down to less than $300 in 1973 dollars. The decision to use a digitally formatted signal has also been vindicated.

VI. Applications

A. Military

The DOD's primary purposes in developing GPS were as follows: 1) its use in precision weapon delivery; and 2) providing a capability that would help reverse the proliferation of navigation systems in the military. Military applications include mine sweeping, aircraft landing, and infantry operations (to name just a few). The Desert Storm campaign was almost a boutique war to demonstrate the effectiveness of GPS. The tactical commanders were finally able to experience the power that comes from precise knowledge of position in a common coordinate frame. It was ironic that the majority of receivers being used were developed by civil companies, with no help from military sponsorship.

B. Dual Use: The Civil Problem

From the beginning of the GPS, it was recognized that the proposed GPS system would provide utility for many more users than the U.S. military. The code structure was arranged to have a precise, protected modulation (the P code), which could be encrypted, and a clear acquisition modulation (the C/A code), which could be exploited by civil users. The earliest presentations always included descriptions of the usefulness to the worldwide civil community. The applications of the GPS for the civilian community are extensive. Initially the GPS was used for accurate time transfer and survey, because these applications could accept the limited initial coverage. Later uses span marine, air, land, and even space. Civil sets currently outnumber military by more than ten to one. That ratio will probably increase as civil set cost decreases. Other chapters discuss in detail many of these applications for these civil receivers.

Table 4 summarizes some of these major civil applications.

This proliferation has led to legitimate fears that the GPS system would be used against its builders, the U.S. Military. Initially it was felt that the P code would demonstrate accuracies about seven times better than C/A.* Therefore, civil receivers would be inherently less accurate. Technology invalidated that assumption. By using carrier aiding, the noise in the C/A receivers could be

*The P code has a chip rate ten times higher, but the C/A code has approximately twice as much power.

Table 4 Some civil applications of GPS

Air navigation	Nonprecision approach and landing
	Domestic en route
	Oceanic en route
	Terminal
	Remote areas
	Helicopter operations
	Aircraft attitude
	Collision avoidance
	Air traffic control
Land navigation	Vehicle monitoring
	Schedule improvement
	Minimal routing
	Law enforcement
Marine navigation	Oceanic
	Coastal
	Harbor/approach
	Inland waterways
Static positioning and timing	Offshore resource exploration
	Hydrographic surveying
	Aids to navigation
	Time transfer
	Land surveying
	Geographical information systems
Space	Launch
	In-flight/orbit
	Reentry/landing
	Attitude measurement
Search and Rescue	Position reporting and monitoring
	Rendezvous
	Coordinated search
	Collision avoidance

smoothed to the point that receiver measurement was an insignificant error source. Anticipating the need to withhold full accuracy from an enemy, the system design had included the ability to degrade the accuracy of the satellite clock or the broadcast satellite location. This so-called selective availability is an important feature for the potential protection of the free world. However, Dr. Parkinson had argued that this capability should not be used all the time, because it could be defeated by various forms of differential (locally corrected) GPS. Degrading the signal continuously would lead to rapid introduction of improved differential techniques. It was felt to be better if S/A were only used when an urgent need was determined.

Alleviating fears of enemy use, any civil differential technique could also be countered in time of hostility by using local area jamming of the more susceptible C/A code. The military would continue to rely on the more jam resistant P code for combat operations.

This civil problem is only partially resolved. The international and civil communities have been pushing for less restrictive civil use. The U.S. government has now agreed to provide the signal worldwide, for the foreseeable future, at the 50 m (one sigma) error level. In addition, they have agreed to give 10 years notice, should they not be able to continue to meet the commitment. Moreover, there is now a joint *military/civil* task force overseeing the operation of GPS with representatives from the departments of defense and transportation.

VII. Pioneers of the GPS

The GPS owes its existence to many foresighted and self-sacrificing people. The following list is not complete; it is hoped that those not mentioned will not feel offended. As is often the case, the engineers who took a concept and made it a reality tend to be forgotten. The writer would like to personally thank the outstanding, and dedicated men and women of the initial JPO who truly made GPS possible with their heroic efforts.

A. Defense Development, Research, and Engineering—Malcolm Currie and David Packard

A staunch and essential supporter from early 1973 was Dr. Malcolm Currie, then Deputy Secretary of Defense for Research and Engineering. In the early years, the GPS suffered because it did not have a single operational Armed Forces command that felt space-based navigation was an operational necessity. Most of the affected commands felt it was desirable, but hoped it would be sponsored (and funded) by someone else. Dr. Currie could visualize the value and threw his support into the bureaucratic fray. Without his intercession, the GPS would have been canceled before the first satellites flew.

Another pioneer—although he may not have been aware of his impact on the GPS program—was David Packard (previously Deputy Secretary of Defense for Research and Engineering). Mr. Packard had brought significant reforms to the DOD decision-making process. This streamlining included brief (7-page) decision coordinating papers (DCP) and crisp decisions after a meeting of the DSARC. Without these reforms, gaining program approval would have been a much longer and more arduous task.

B. Commander of SAMSO, General Ken Schultz

As Program Director, Dr. Parkinson was extremely fortunate to work for a general who also had been a Program Director. Lt. General Kenneth Schultz was tough and fair and knew how difficult it was to run a large program. He laid down the objectives but did not attempt to control the process totally. Along the way, he taught the JPO many essential things about keeping a program on track, from procurement to personnel.

C. Contractors

The principal Phase I hardware contractors are listed in Table 5. Literally hundreds of hardworking, capable engineers and managers produced the Phase I success.

Table 5 Principal contractors

Contractor	Development effort	Comments
Rockwell International	Development satellites	Initially three plus a flying Qual model; two more added later
General Dynamics	Control segment and direction of Magnavox	Also developed the inverted range for testing
Magnavox	User equipment	Included monitor receivers for the control segment
Texas Instruments	User equipment	An alternate, competitive receiver source
Collins Radio of Rockwell International	Jam resistant user equipment	Actually under contract to the U.S. Air Force Flight Dynamics Lab
Stanford Telecommunications	Signal structure	Also instrumental in obtaining extra satellites

D. Joint Program Office Development Team

It was members of this team who synthesized the design of GPS in 1973, prior to signing contracts with any of the support contractors. From the beginning, a conscious decision was made *not* to use an integrating contractor. The overall integration was to be handled by the Joint Program Office in cooperation with various contractors. For this to work, the JPO had to manage the technical tradeoffs and all major systems interfaces.

Fortunately, Dr. Parkinson had been given strong support in directing the effort. This included careful selection of the Air Force officers who had to make the management and technical approach succeed. Of the approximately 35 military officers involved, 6 held Ph.D.s in engineering, and virtually all the others held Master's degrees. Many had experience in running large programs, and some (who had been at the Air Force's Central Inertial Guidance Test Facility) were extremely skilled in devising and running test programs for navigation systems. With the extensive travel demands, it was essential that Dr. Parkinson had a strong and effective deputy to smooth administration of the complex program. That role was filled effectively by then Lt. Col. Steve Gilbert during the initial phase of development and later by Lt. Col. Don Henderson. In addition, the JPO was supported by a small, but effective, cadre of engineers from the Aerospace Corp., which was initially headed by Walt Melton, who had done much of the pioneering work on 621B.

E. Predecessors

The Air Force 621B program had been developed in the Plans Directorate of the Air Force's Space and Missile Systems Organization with strong support from the Aerospace Corp. Two individuals who pioneered the most essential pieces of other GPS technology were Roger Easton of the NRL, who developed the initial clock technology, and Dr. Richard Kirschner of the ARL, who had developed Transit.

VIII. Future

The quiet revolution of NAVSTAR GPS has just begun. Given that the number of active satellites in the constellation now has reached 24, the use of GPS surely will expand. As that expansion continues, the demand will be for greater integrity, which will lead to a modest increase in the number of satellites. The major issues awaiting resolution are the following:

1) Sufficiency of the number of satellites

2) Expansion of the backup control segment capability

3) Resolution of the international request for some civilian control of the system

An expanded dialogue between the military operators and civilian users has begun. The next great wave of progress will be *differential* GPS systems, which squeeze the expected dynamic errors down to less than 1 m. All users will benefit as this new "navigation utility" comes into full operation and usefulness over the next 20 years.

References

[1]Parkinson, B. W., "Overview," *Global Positioning System,* Vol. I, Institute of Navigation, Washington, DC, 1980, p. 1.

[2]Easton, R. L., "The Navigation Technology Program," *Global Positioning System,* Vol. I, Institute of Navigation, Washington, DC, 1980, pp. 15–20.

[3]Bowen, R. et al., "GPS Control System Accuracies," *Global Positioning System,* Vol. III, Institute of Navigation, Washington, DC, 1986, p. 250.

[4]Sharrett, Wysocki, Freeland, Brown, and Netherland, "GPS Performance: An Initial Assessment," *Proceedings of ION GPS-91,* Institute of Navigation, Washington, DC, 1991.

[5]Anon., *Proceedings of ION-89,* Institute of Navigation, Washington, DC, 1989, p. 19.

Chapter 2

Overview of GPS Operation and Design

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089
and
Bradford W. Parkinson†
Stanford University, Stanford, California 94305

I. Introduction to GPS

THE Global Positioning System (GPS) consists of three segments: the space segment, the control segment, and the user segment, as shown in Fig. 1. The *control segment* tracks each satellite and periodically uploads to the satellite its prediction of future satellite positions and satellite clock time corrections. These predictions are then continuously transmitted by the satellite to the user as a part of the navigation message. The *space segment* consists of 24 satellites, each of which continuously transmits a ranging signal that includes the navigation message stating current position and time correction. The *user receiver* tracks the ranging signals of selected satellites and calculates three-dimensional position and local time.

This chapter is designed to provide a summary discussion of the GPS. Later chapters develop the details. All segments of the system, along with a detailed discussion of the signal and the multiple applications of the GPS, are covered in separate chapters.

II. Performance Objectives and Quantitative Requirements on the GPS Signal

The key performance objectives of the GPS system can be summarized as follows:

1) High-accuracy, real-time position, velocity, and time for *military users* on a variety of platforms, some of which have high dynamics; e.g., a high-performance

Copyright © 1994 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Released to AIAA to publish in all forms.

*Ph.D., Chairman of the Board.

†Professor, Department of Aeronautics and Astronautics, and Director, GPS Program.

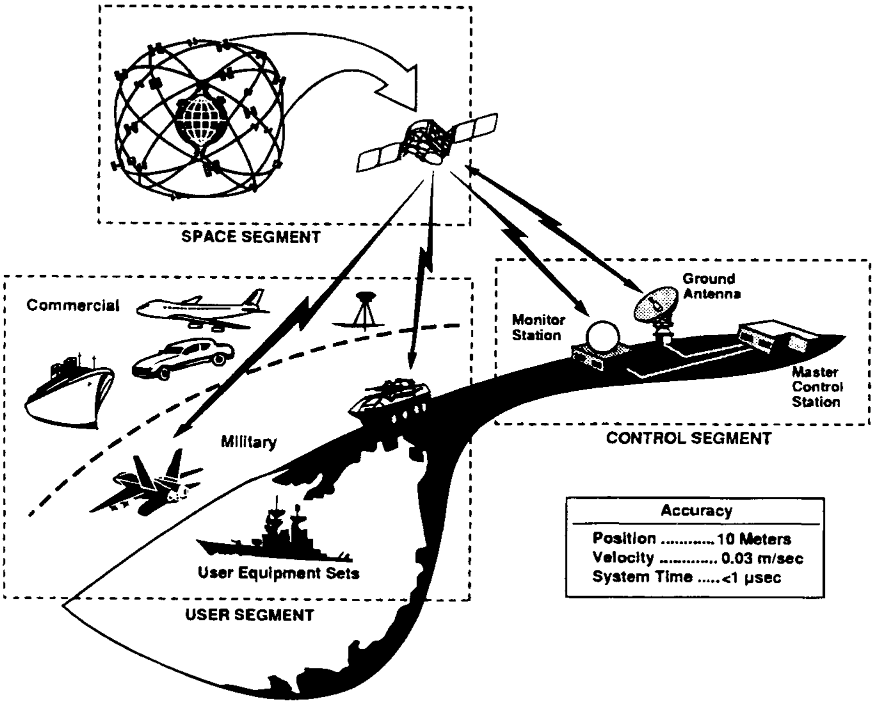


Fig. 1 GPS consists of three segments: space, control, and user; the user segment contains both civil and military users.

aircraft—high accuracy translates into 10-m three-dimensional rms position accuracy or better; velocity accuracy < 0.1 m/s.

2) Good accuracy for *civil users*—the real-time civil user accuracy objective is considered to be 100 m (at about the 95th percentile) or better in three dimensions. In the future, this accuracy may be improved by reducing or eliminating the deliberate degradation of the ranging signal.

3) Worldwide, all weather operation, 24 h a day.

4) Resistance to intentional (jamming) or unintentional interference for all users—enhanced resistance to jamming for military users.

5) Capability for highly accurate geodetic survey to centimeter levels using radio frequency carrier measurements—capability for high-accuracy time transfer to 100 ns or better.

6) Affordable, reliable user equipment—users cannot be required to carry high-accuracy clocks; e.g., atomic frequency standards, or sophisticated arrays of directional antennas that must be pointed at the satellites.

In addition to these performance requirements for the user, the GPS must also employ a cost-efficient space segment, must live within constraints of available bandwidth and frequency allocations, and have a control segment capable of measuring the satellite orbits, clocks, and uploading data to the spacecraft for retransmission to the users.

This chapter provides a summary of the GPS and describes how these accuracy and other performance requirements translate into requirements on the GPS signal. For example, user position accuracy translates into accuracy requirements for the time measurement performed by the GPS user receiver. These requirements also affect the radio frequency frequency selection. Because the satellites must be limited in transmit power, and there are many perturbing physical phenomena and geometrical considerations as discussed later, all of these selections have been made with care.

A. Satellite Navigation Concepts, Position Accuracy, and Requirement Signal Time Estimate Accuracy

As an elementary step in discussing the use of satellites for real-time navigation, consider the single idealized navigation satellite and a single user, as shown in Fig. 2. Assume that the user is fixed in inertial space on a nonrotating Earth. Assume also that the satellite has information as to its precise position vs. time and contains a perfect clock. Imagine that both position and time are displayed in lights on the side of the spacecraft so that they are observable to the user who has a telescope and camera. For this example of the principal of satellite ranging, the satellite and user coordinates are both expressed in an Earth–centered-inertial (ECI) nonrotating coordinate system with the origin at the Earth's center (see Fig. 3) and we neglect atmospheric and relativistic effects.

In this example, the user camera periodically photographs the satellite clock and position indicator and compares the satellite clock reading t_s with a simultaneous

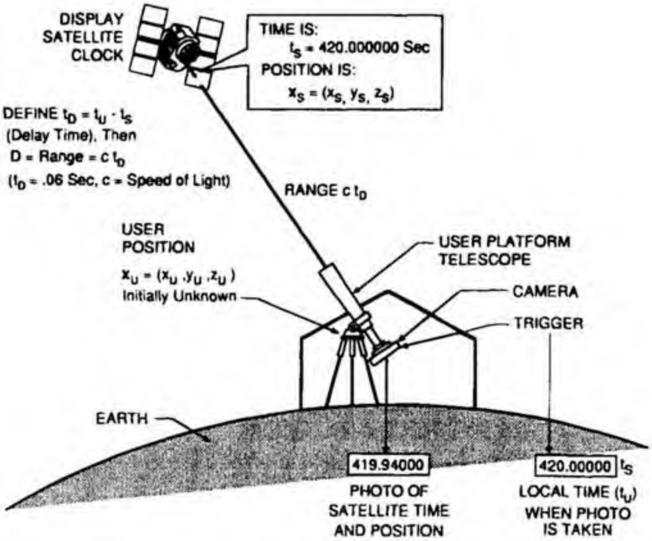

DISPLAY SATELLITE CLOCK
TIME IS: t_S = 420.000000 Sec
POSITION IS: $x_S = (x_S, y_S, z_S)$
DEFINE $t_D = t_U - t_S$ (Delay Time). Then
D = Range = c t_D
(t_D = .06 Sec, c = Speed of Light)
RANGE c t_D
USER POSITION
$x_U = (x_U, y_U, z_U)$
Initially Unknown
USER PLATFORM TELESCOPE
CAMERA
TRIGGER
EARTH
419.94000
PHOTO OF SATELLITE TIME AND POSITION
420.00000 t_S
LOCAL TIME (t_U) WHEN PHOTO IS TAKEN

Fig. 2 Satellite and user clock timing concepts—photographs of the satellite clock are taken by the user. Coordinates are expressed in an Earth–centered-inertial (ECI) coordinate system. The true system time is t_u at the time of reception. A perfect user clock is assumed. The satellite position can be alternately denoted by x_s in Cartesian coordinates or r_s in radius vector coordinates.

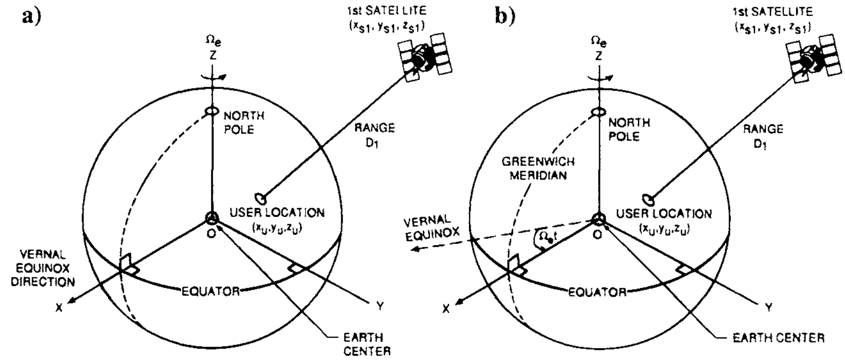


Fig. 3 Earth–centered-inertial and Earth-centered, Earth-fixed (ECEF) coordinates: a) the ECI coordinates are nonrotating with the x-axis aligned with a vector from the Sun to the Earth position at the vernal equinox (the first day of autumn); b) ECEF coordinates rotate with the Earth with the x-axis on the Greenwich meridian.

reading of the local user clock t_u. Also for this example, both clocks are assumed to be exactly synchronized to system time, and the user is assumed to be stationary. The photograph reveals that, at the receive time instant when the user clock shows $t_u = 420.00000$ s, the image received at the camera from the satellite showed the satellite position at $\boldsymbol{x}_s(t_s)$ [or $\boldsymbol{r}_s(t_s)$ in radius vector coordinates] and $t_s = 419.94000$; i.e., it shows the satellite clock and position 0.060 s earlier. Thus, there is a measured delay caused by the finite speed of light c of $t_D = t_u - t_s = 0.06$ s and a range to the satellite $D = c\ t_D$ measured at time $t = t_u$. Therefore, the receiver's location at time t_u is somewhere on a sphere of radius D centered at $\boldsymbol{x}_s(t_s)$. This simple example does not address the impact of the satellite's velocity nor the possibility of a dynamic user. Note also that in order to convert this result to meaningful user coordinates; e.g., Earth-centered, Earth-fixed (ECEF) coordinates, we must account for the fact that the Earth is rotating through 360 deg in inertial space per sidereal day. (A sidereal day is approximately 23 h 56 min 4 s of mean solar time.)

Clearly, if we were able to perform the same type of measurement with three satellites simultaneously, we could locate the user position in three dimensions at the intersection of three spheres and perform the desired real-time navigation. However, doing so requires accurate, synchronized time at the user terminal. Assume now that the user clock has an unknown bias error b_u and thus the user clock reads $t_u' = t_u + b_u$, where t_u is the "true" system time at the user's time of reception. By adding a ranging measurement to a fourth satellite, the solution can be found for both $\boldsymbol{x}(t_u)$ and true user clock time t_u (or user clock error Δt_u), at the time at which the measurement is taken (see Fig. 4).

The difference between satellite clock time and user clock time when the user clock is not precise is termed "*pseudorange.*" For the ith satellite, this range difference is denoted as ρ_{iT} where subscript T denotes true pseudorange. The *true pseudorange* ρ_{iT} to satellite i, in the idealized error-free condition, is the true range plus the user clock bias correction b_u expressed in seconds and is expressed as $\rho_{iT} = c(t_u - t_{si}) + c\ b_u$. However, in the real measurement, there are random

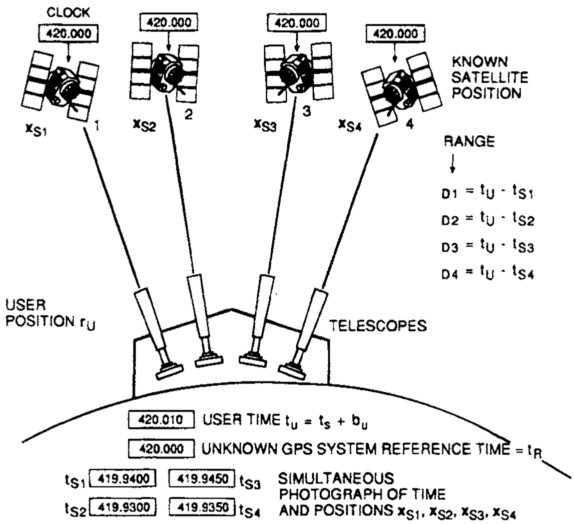


Fig. 4 Three-dimensional user position and clock bias measurement for a user with an accurate user clock. Four pseudorange equations are needed to solve for the four unknowns, user *x*, *y*, and *z* coordinates and user clock bias b_u. Note that the satellite positions are all observed by the user at slightly different times t_{si}.

noise effects, various other bias errors, propagation errors (and relativistic effects) so that the *measured pseudorange* ρ_i is $\rho_i = \rho_{iT} + \Delta D_i - c\ \Delta b_i + c\ (\Delta T_i + \Delta I_i + \nu_i + \Delta \nu_i)$ where Δb_i is the satellite bias clock error (s); ΔD_i is the satellite position error effect on range; ν_i is the receiver measurement noise error for satellite i (s); ΔI_i is the ionospheric delay error (s); ΔT_i is the tropospheric delay error (s); and $\Delta \nu_i$ is the relativistic time correction (s). Thus, the actual user clock reads $t'_u = t_u + b_u$, and the actual clock of satellite i reads $t'_{si} = t_{si} + \Delta b_i$, where t_u and t_{si} are the true system times at the user at the time of reception and at satellite i at the time of its signal transmission, respectively.

We define *true pseudorange* $\rho_{iT} = |\boldsymbol{x}_{si} - \boldsymbol{x}_u| + cb_u = |\boldsymbol{r}_{si} - \boldsymbol{r}_u| + cb_u = D_i + cb_u$. Note that pseudorange defined in this manner is not directly an observable, and each of these quantities may vary with time. Note also that the techniques discussed here permit real-time satellite navigation and position measurement. This technique is distinct from systems that must observe the changes in satellite Doppler shift over some period of time as the satellite passes overhead to determine position. Techniques of this latter type clearly cannot autonomously determine the position of a rapidly moving vehicle.

Although the preceding examples of Figs. 2 and 4 used optical measurements, exactly the same process can be performed using microwave signaling with coded signals. Examination of the fine detail of the received signal code provides exactly the same information as a photograph of a clock, because the structure of the code can be interpreted as time counts of a clock. One simple radio frequency analogy to the optical clock display is a signal that is modulated by a 10-stage binary counter that counts the number of precisely timed 1 μs clock cycles up to $2^{10} = 1024$ and then repeats. Examination of the last 10 bits of the binary waveform then gives coarse time analogous to hours and minutes, and examination of the exact timing of the clock transition gives the finer resolution of time. As compared to a system that simply monitors periodic sequences of identical 1 μs pulses, the counter, in effect, reduces the level of ambiguity in the measurement from 1 to 1024 μs, etc.

The preceding simplified description of the GPS positioning calculations briefly introduces a number of significant effects, which are treated in detail in later chapters of this volume. These include the following:

1) *User motion*—The user is generally moving in inertial space, and we must account for the user motion between the time the signal is transmitted from the satellite and the time when the signal is received.

2) *Atmospheric effects*—Excess delay caused by the wave traveling through the atmosphere (troposphere and ionosphere) must be estimated or measured.

3) *Relativistic effects*—There are a number of effects caused by satellite and user motion, Earth rotation, and the Earth's gravitational field, all of which can be significant.

1. User Receiver Measurements—Pseudorange and Carrier Phase

Next we examine how a receiver makes the measurements on the signal waveform or the radio frequency carrier to form an accurate estimate of the pseudorange ρ_i. Figure 5 illustrates some of the range measurements that can be

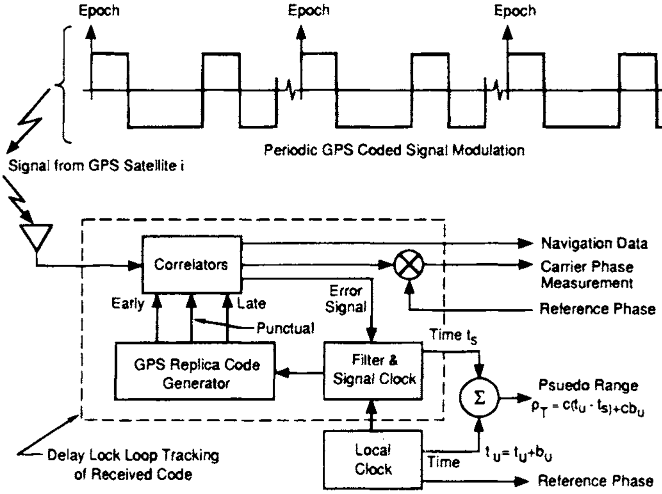

Epoch
Epoch
Epoch
t
Periodic GPS Coded Signal Modulation
Signal from GPS Satellite i
Correlators
Navigation Data
Carrier Phase Measurement
Reference Phase
Early
Late
Punctual
Error Signal
Time t_S
GPS Replica Code Generator
Filter & Signal Clock
Σ
Psuedo Range
$\rho_T = c(t_U - t_S) + cb_U$
Delay Lock Loop Tracking of Received Code
Local Clock
Time
$t_U = t_U + b_U$
Reference Phase

Fig. 5 Pseudorange measurement using a delay–lock-loop (DLL). The GPS replica code generates early and late reference signals that are both fed to a correlation device that produces an estimate of whether the early or late signal provides the best match. The resulting error signal is then used to control the signal clock in a tracking mode. The differences between the two clocks t_u'–t_s is then multiplied by c to form pseudorange $c(t_u' - t_s)$. Similar measurements are made on the received carrier phase. The user receiver has a clock bias offset b_u in seconds. (Satellite clock error is ignored here.)

made on a signal waveform. Pulses of radio frequency energy can be modulated with a special pulse code sequence that has a distinct beginning (or epoch). The GPS civilian [coarse acquisition (C/A)] signal repeats such a sequence every millisecond. The C/A code is a random-like or pseudonoise (PN) binary sequence of 1023 chips. The pseudorange $t_u - t_s$ can then be recovered in a special delay lock loop receiver designed to track and to detect the code from that satellite, as shown in Fig. 5. The measurement relies on the unique code properties of each satellite signal. These properties enable the receiver to measure pseudorange to each satellite separately. The delay lock technique creates an internal replica of the known modulation sequence and adjusts the internal epoch until it exactly matches the received signal in delay. This matching is performed by cross-correlating the received and internal signals and finding the start time that maximizes the correlator output. The satellite clock time at the time of transmission is then subtracted from the user clock time to recover the measured pseudorange.

If the user receiver clock t_U and the satellite clock are both synchronized to GPS reference time,* then range $= D = c(t_u - t_s)$. For purposes of this discussion, atmospheric and other propagation path delay perturbations are neglected. Pseudorange is the same measurement when the user receiver clock t_U has an unknown and possibly time varying clock bias b_u expressed in seconds. The pseudorange is then $\rho = D + cb_u$, as shown in Fig. 5. This technique is known as *code ranging* to distinguish it from carrier phase measurements.

Most GPS receivers can also reconstruct the GPS radio frequency carrier at 1575.42 MHz and use this sinusoid as a ranging signal. This measurement is very precise (typically subcentimeter or a fraction of the 19-cm wavelength), but its *accuracy* is limited by the difficulty of resolving which cycle is being received (called the cycle ambiguity or $n\lambda$ problem). Carrier phase accuracy corresponds to the equivalent carrier phase noise expressed in distance which would be on the order of 2 mm if the carrier phase could be measured to 1% of the wavelength. However, the initial value of the carrier phase is completely ambiguous, and we must resort to the use of various phase differencing techniques. Nonetheless, it is possible to measure changes in phase both very precisely and without significant cycle slipping over many seconds. This carrier phase measurement can then allow us to make very precise position measurements. The use of carrier phase measurements for various surveying and aircraft landing applications is discussed in later chapters in this volume. Table 1 gives a rough estimate of the measurement accuracy, bias, and precision for GPS carrier and code measurements.

a. Relating pseudorange accuracy and positioning accuracy—dilution of preci.ion (DOP). Positioning accuracy reflects the final capability of most GPS receivers. Although it is related to *ranging* accuracy, they are not the same. The relationship between them is a function of the geometry of the selected satellites; that is, the directions of arrival of the satellite signals. To achieve a positioning accuracy requirement of 10 m, the ranging accuracy and geometry must both combine to acceptable values. For example, if each individual pseudorange measurement has a statistically independent error of zero mean with the same rms

*All time intervals can be expressed in equivalent distance in meters by multiplying by the speed of light c.

Table 1 Rough order of magnitude estimates of GPS code and carrier phase measurements in meters[a]

	Rough measurement accuracy	
Measurement	Bias	Precision
Code	≈ 5 m	≈ 2 m
Carrier phase	$n\lambda$	≈ 0.002 m

[a] There is assumed to be no selective availability (SA) degradation. The value of n in the carrier phase ambiguity must be determined.

value of σ (caused by all effects), then the rms position errors are given by the following:

Position error = DOP $*$ σ, where DOP is a multiplier determined by the geometry and is typically between 1 and 100.

The quantity DOP is calculated from the unit vectors to each of the satellites, as shown in Chapters 5 and 11, this volume. Generally, if the DOP rises above six, the satellite geometry is not very good. There are several measures of positioning accuracy. For the current nominal constellation, the worldwide median position dilution of precision (PDOP) (50th percentile) is approximately 2.5. As another example, the horizontal error (in the x and y coordinates) is given by rms error horizontal = HDOP (horizontal dilution of precision) $*$ σ, where rms denotes the root mean square error. The vertical error (in the z coordinate) is found from rms error vertical = VDOP (vertical dilution of precision) $*$ σ.

The speed of light is approximately $< \cong 3$ m/ns and l/c $\cong$ 3.3 ns/m. If the satellite geometry produces an HDOP of 3 (and a horizontal error less than 10 m is required), then HDOP $* \, \sigma = 3\sigma \leq 10$ m $*$ 3.3 ns/m. Thus, the required ranging accuracy is then $\sigma \leq 11$ ns. In a similar manner, the rms position error (in three dimensions x, y, and z) is estimated by the relationship rms position error = PDOP $*$ σ.

Table 2 summarizes some of the key definitions of parameters in the GPS user position calculations and error sources.

III. GPS Space Segment

A. GPS Orbit Configuration and Multiple Access

The discussions of the basic concepts for GPS and analysis of the geometric dilution of precision have shown that the user must make measurements on four or more satellites simultaneously to provide real-time three-dimensional navigation. Thus, the satellite orbital constellation must provide a user anywhere in the world simultaneous access to measure pseudorange to four or more satellites at any time, 24 h a day. Furthermore, as shown later in the DOP analysis, the satellites should be widely spaced in angle. Measurements on each of the four or more satellites must be made simultaneously or nearly simultaneously without mutual interference if we are to solve for position. This capability is termed

Table 2 Summary of notation for GPS position and pseudorange parameters

Parameter	Description
$\mathbf{x}_{si}$ or $\mathbf{r}_{si}$	Position of satellite i in either x, y, z or radius vector coordinates, respectively. The position of the satellites at the time of transmission is $\mathbf{x}_{si}$ (t_{si}) or $\mathbf{r}_{si}$ (t_{si}).
$\mathbf{x}_u$ or $\mathbf{r}_u$	User position in either x, y, z or radius vector coordinates. User position may also vary with time. User position at time of reception is $\mathbf{x}_{ui}$ (t_{Ri}) or $\mathbf{r}_{ui}$ (t_{Ri}).
t_{si}	True time of transmission from satellite i. This parameter is the true time of transmission and may not be exactly the same as indicated by the satellite.
t'_{si}	Actual satellite clock reading $t'_{si} = t_{si} + \Delta b_i$
Δb_i	Satellite bias clock error (expressed in s)
t	GPS system time
t'_u	Actual user clock time at time of reception of signal $t'_u = t_u + b_u$
t_u	True user time at time of reception
b_u	User clock bias—can vary with time
ρ_{iT}	True value of pseudorange $\rho_{iT} = \lvert\mathbf{x}_{si} - \mathbf{x}_u\rvert + b_u = \lvert\mathbf{r}_{si} - \mathbf{r}_u\rvert + b_u$
ρ_i	Measured pseudorange with various error contributors
$n(t)$	Receiver thermal noise
Δb_i	Satellite bias clock error (expressed in s)
ΔD_i	Satellite position bias error effect on range
ν_i	Receiver pseudorange measurement noise error for satellite i, s
ΔI_i	Ionospheric delay (expressed in s)
ΔT_i	Tropospheric delay (expressed in s)
$\Delta \nu_i$	Relativistic time correction (expressed in s)
c	Velocity of light
DOP	Dilution of precession
VDOP, HDOP, PDOP	Vertical, horizontal, and position geometric dilutions of precession
GDOP	Geometric dilution of precession includes both position and time error effects
ϕ_i	Measured carrier phase offset as received from satellite i
ϕ_{iT}	True received carrier phase offset $\phi_{iT} = \omega_o(t_{si} - t_u)$ at frequency ω_o
λ	Wavelength of GPS carriers can have value λ_{L_1} or λ_{L_2} for GPS L_1 or L_2 frequencies.
n	Carrier phase cycle count ambiguity
D_i	True range from satellite i to user $\lvert\mathbf{x}_{si} - \mathbf{x}_u\rvert = \lvert\mathbf{r}_{si} - \mathbf{r}_u\rvert = c(t_u - t_{si})$

multiple access. Multiple access signaling permits measurements to be made on a signal from one satellite without signals from other satellites interfering with that measurement.

From a user performance standpoint, satellite orbit altitude selection has several effects:

1) The higher the orbit altitude, the greater the fraction of the Earth visible by a single satellite

2) Within limits, power flux density on the Earth is nearly independent of orbit altitude because the satellite antenna beamwidth can be selected (widened or narrowed) to provide full Earth coverage.

3) A low-orbit altitude with its corresponding short visibility time leads to a larger number of signal acquisitions and satellite–satellite handovers by the user receiver, and larger Doppler shift must be tolerated in the receiver.

The selected satellite orbital constellation contains 24 satellites, the GPS-24, in six orbit planes. There are four satellites in each of the six planes, as shown in Fig. 6. The satellites have a period of 12 hours sidereal time* and a semimajor axis of 26,561.75 km. A satellite with an orbit period of 12-h sidereal time produces a ground track (projection on the Earth's surface) which repeats over and over. For GPS the longitude crossing at the equator is kept fixed to within ± 2° by the GPS Control Segment. This orbit has 63% of the radius of a geostationary satellite orbit with a 24-h period. The GPS semimajor axis is the orbit radius of the circular GPS orbits, and thus the GPS satellites have an altitude of 20,162.61 km above the Earth's equatorial radius[1] of 6378.137 km. The altitude of the GPS orbit obviously is well above the atmosphere and not subject to atmospheric drag. Other perturbations such as solar pressure and lunar and solar gravitational orbit perturbations can be significant. The satellites are inclined with respect to the equator by 55 deg (the initial GPS satellites had a 63-deg inclination). Table 3 summarizes the approximate parameters of the GPS orbit. The satellite orbital constellation is described in detail in Chapter 5, this volume.

B. GPS Satellite Payload

The key role of the satellites is to transmit precisely timed GPS signals at two *L*-band frequencies† 1.57542 GHz and 1.2276 GHz. These signals must have embedded in them, in the form of navigation data, both the precise satellite clock time as well as satellite position so that a user receiver can determine both satellite time t_{si} and satellite position $\boldsymbol{r}_{si}$ at the time of transmission. These navigation data are uploaded from the GPS control segment (CS) to each satellite and then stored in memory in the satellite for readout in the satellite navigation data stream. Figure 7 shows a simplified view of the GPS satellite payload. The GPS upload station sends the satellite the ephemeris information regarding the satellite orbit

*A sidereal day is defined as the time for the Earth to complete one revolution on its axis in ECI space and consists of 24 mean sidereal hours where 1 mean sidereal day is slightly shorter than a mean solar day. One sidereal day is 23 h, 56 min, 4.009054 s or 86,164.09054 s of mean solar time. One mean sidereal day is equal to 0.997269566 mean solar day or one mean solar day is equal to 1.002737909 mean sidereal day.

†These signals, as well as the L_3 signal at 1381.05 MHz, are all selected and filtered so as to minimize interference with the radio astronomy bands (see Chapter 3, this volume).

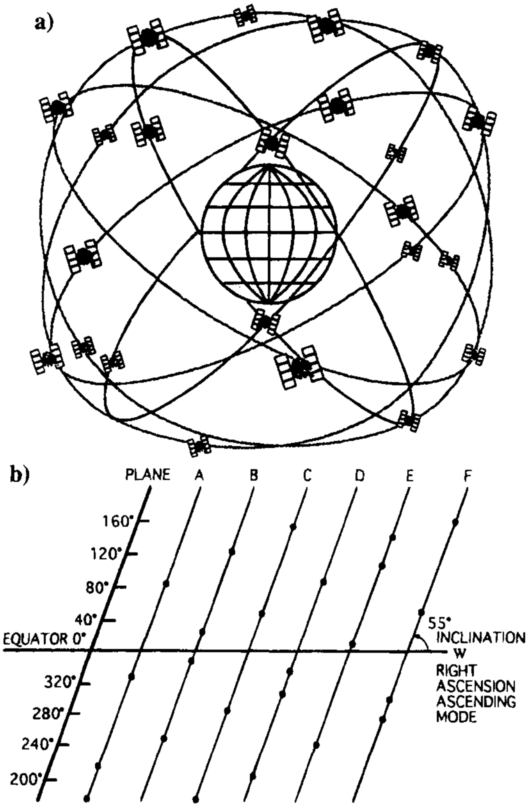


Fig. 6 GPS satellite constellation: a) the six orbit planes shown in ECI coordinates; and b) satellite positions on each of the six orbit planes. The GPS constellation has satellites in six equally spaced orbit planes. The present GPS-24 satellites constellation shown in b) contains 24 satellites. The uneven satellite phasing in each plane is designed to minimize the effect of satellite outage.

and the exact position in that orbit vs. time. Included is a satellite clock correction that calibrates the offset of the satellite clock relative to GPS system time. These data are uploaded to the satellite through an S-band telemetry and command system.

One of the keys to GPS satellite performance is the stability of the GPS satellite clocks. Each satellite carries redundant atomic oscillators of high stability. These atomic clocks are stabilized using either rubidium or cesium atoms in gaseous form.

The atomic clocks along with appropriate frequency synthesizers then synchronize the GPS signal generators and also control the radio frequency center frequencies of the two *L*-band frequencies. The signals are then amplified and filtered to remove signal power outside the allocated frequency bandwidth of 20 MHz for each of the two *L*-band signals. The signals also are modulated by the navigation data that carry the satellite position and time information to the user.

Table 3 Approximate GPS satellite parameters

Orbit plane	Six equally spaced ascending nodes at 120 deg
Orbit radius r_{cs}	26,561.75 km semimajor axis
Orbit velocity (circular) (ECI)	$= \sqrt{\frac{\mu}{r_{cs}}} = 3.8704$ km/s
Eccentricity	Nominally zero, but generally less than $e = 0.02$
ω_s angular velocity	1.454×10^{-4} rad/s
Period[a]	12 h mean sidereal time
Inclination	$i = 55$ deg nominal

[a] The period of an orbit in seconds of mean solar time is $T_p = (2\pi/\sqrt{\mu})a^{3/2}$ where a is the semimajor axis in meters and μ is the Earth's gravitational parameter $\mu = 3.986005 \times 10^{14}$ m^3/s^2. For 12-h mean sidereal time period, $a = 26561.75$ km including a minor correction for a nonspherical Earth. A new model, Joint Gravity Model #2, is being proposed by NASA and the University of Texas, which gives $\mu = 3.986004415 \times 10^{14}$ m^3/s^2, and $R_e = 6378.1363$ km for the Earth's mean equatorial radius.

(The reader is referred to Chapter 6, this volume on the GPS satellite payloads for a more thorough description.) One of the limitations on GPS clock accuracy is selective availability. Selective availability is a clock dither that can be imposed on the GPS signals to restrict unauthorized (nonmilitary) access to the full accuracy of the system.* Chapter 1 of the companion volume describes differential GPS that can be used to improve civil accuracy.

C. Augmentation of GPS

As discussed in later chapters in this volume and the companion volume, we can augment the GPS satellite system with other ranging signal sources:

1) Ground transmitters or pseudolites that transmit GPS signals and other information to support GPS—a special form is the integrity beacon used for aircraft landing

2) Additional satellites can either carry transponders that can relay GPS-type signals from synchronized ground transmitter uplinks or have navigation payloads similar to the basic GPS satellites. The Federal Aviation Administraton's Wide Area Augmentation System (WAAS) is an example of an augmentation of GPS. This system adds geostationary relay satellites.

3) Differential and wide area differential ground stations. These ground stations transmit correction information to appropriately equipped GPS receivers to improve the accuracy of the receivers.

IV. GPS Control Segment

The GPS CS has several objectives:

1) Maintain each of the satellites in its proper orbit through infrequent small commanded maneuvers.

*It is expected that selective availability effects will eventually be eliminated. A recent report of the National Research Council[5] has recommended that selective availability be turned to zero.

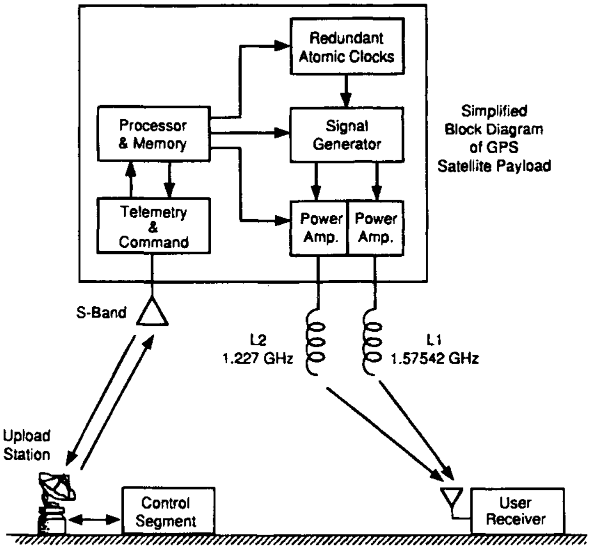


Fig. 7 Simplified GPS satellite payload functional diagram.

2) Make corrections and adjustments to the satellite clocks and payload as needed.

3) Track the GPS satellites and generate and upload the navigation data to each of the GPS satellites.

4) Command major relocations in the event of satellite failure to minimize the impact.

Although each of these objectives is important, this discussion concentrates on the third objective.

The Operational Control Segment began operation in 1985 and consists of five monitor stations, four ground antenna upload stations, and the Operational Control Center. Each of these facilities is shown in Fig. 8. The sites have been selected to provide a significant separation in longitude between each of the monitor stations. Each of these sites, except the site in Hawaii, also contains a ground antenna upload station.

A. Monitor Stations and Ground Antennas

Each of the five monitor stations contains multiple GPS tracking receivers designed to track both the L_1 and L_2 codes and carriers for each of the satellites in view. The monitor stations also contain redundant cesium standard clocks for the GPS receivers to use as a reference oscillator and also to time tag each of the measurements. The measurements of code clock delay and carrier phase for each satellite in view are then sampled, time tagged, and multiplexed in a datastream to send back to the operational control center. Each of the four ground antenna (GA) upload stations has the capability of uploading navigation data to the satellites on an S-band T T&C link. As discussed earlier, the visibility region around each GA extends approximately ±72 deg in Earth angle about the GA.

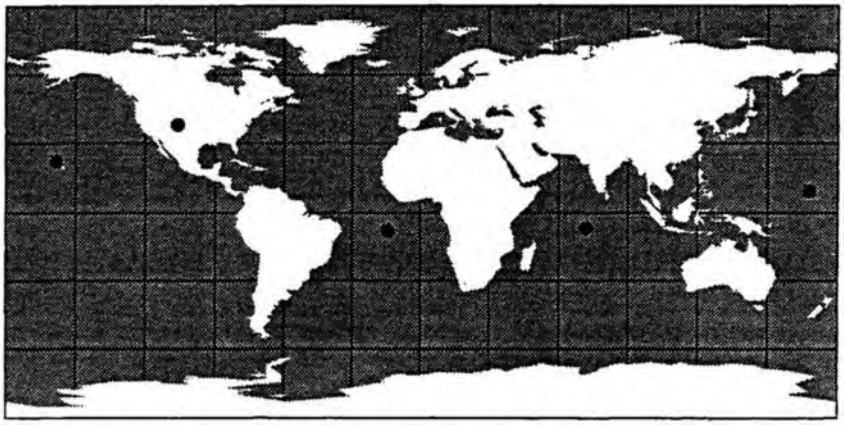


Fig. 8 GPS control segment. There are monitor stations at Hawaii, Colorado Springs, Ascension Island, Diego Garcia in the Indian Ocean, and at Kwajalein Island in the West Pacific.

B. Operational Control Center

The operational control center receives the multiplexed pseudorange measurements and carrier measurements from each satellite in the L_1 and L_2 carriers. The Kalman filter processor in the OCS then estimates the ephemerides, clock error, and other navigation data parameters; e.g., satellite health, for each satellite. The objective of the OCS is to format navigation data for a minimum of 14 days of updates. Navigation data are then transmitted to the upload ground antennas. Each satellite can be given a fresh upload three times a day, approximately eight hours apart. However, normally there is only one upload per day. (See Chapter 16, this volume.) Each upload contains many pages of navigation data that are then fed to the GPS spacecraft processor. The GPS satellite processor then reads out the appropriate set of navigation data for the specific time period appropriate to the time of transmission.

Satellite clock errors in the navigation data are the dominant source of user range error (URE) when the time since the last upload reaches several hours.* Even with three uploads per day, computer simulations run in 1985 show that satellite clock predictability based on clock specification values can limit the GPS user range accuracy to 3m (1σ) 10 h after the satellite ephemeris and clock prediction upload using the specified Allan variance for the satellite clock (2×10^{-13} at $\tau \geq 61{,}200$ s and 7×10^{-12} at $\tau \leq 50$ s).[2] However, actual satellite clocks perform better than the specification. The specified URE is computed as the following rms sum: URE = $[(\text{radial perturbation})^2 + 0.0192\ (\text{in-track perturbation})^2 + 0.0192\ (\text{crosstalk perturbation})^2]^{1/2}$, which is a representative error projection to the user.

The simulations showed the total URE had an rms value for 10-h updates of 4.2m (1σ), which was dominated by the clock error component for the specified

*This statement assumes that both L_1 and L_2 are available. If not, ionospheric modeling errors are often the largest.

clock and is well within the specified 6m maximum. By way of comparison, the simulated URE for 0 h, 3 h, 6 h, and 24 h predictions were 1.4m, 2.4m, 3.2m, and 8.4m (1σ), respectively. Most (95%) of the URE at 10 h is a result of the accumulated clock noise. When the simulated clock stability was reduced to the typical observed value (rather than specified levels) of 1×10^{-13}, the URE at 10 h decreased to 2.3m (1σ). For further information on expected errors see Chapters 11 and 16, this volume. A discussion of the GPS control segment is found in Chapter 10, this volume.

V. GPS User Segment

There are a great many applications for the GPS system. New applications seem bounded only by the imagination. This section lists some of the more common modes of operation. All of these are discussed in more detail elsewhere in this volume. Although the primary purpose of GPS was a military application, civil users are already more prevalent. The section begins however by describing the fundamental user system architecture.

A. GPS User Receiver Architecture

A generalized view of a typical GPS user system is depicted in Fig. 9. This section discusses the basic configuration of each of these elements. Although most receivers employ only a single GPS antenna, the generalized GPS receiver begins with one or more antenna/low noise amplifiers. More than one antenna/amplifier may be employed in order to achieve the following:

1) Accommodate maneuvering of the user platform; e.g., an aircraft banking

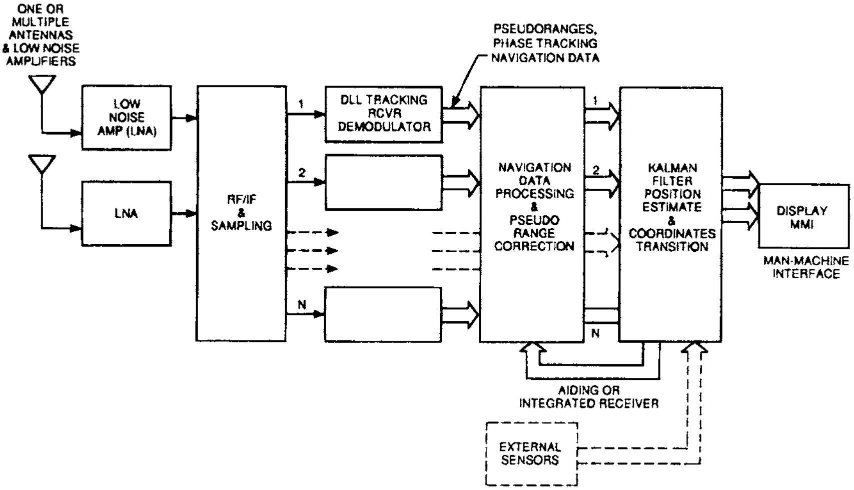

ONE OR MULTIPLE ANTENNAS & LOW NOISE AMPLIFIERS
LOW NOISE AMP (LNA)
LNA
RF/IF & SAMPLING
1
2
N
DLL TRACKING RCVR DEMODULATOR
PSEUDORANGES, PHASE TRACKING NAVIGATION DATA
NAVIGATION DATA PROCESSING & PSEUDO RANGE CORRECTION
1
2
N
KALMAN FILTER POSITION ESTIMATE & COORDINATES TRANSITION
DISPLAY MMI
MAN-MACHINE INTERFACE
AIDING OR INTEGRATED RECEIVER
EXTERNAL SENSORS

Fig. 9 Generalized GPS user system configuration with separate receiver and position estimating functions.

and thereby avoid blocking some of the satellites with a wing

2) Increase the antenna gain

3) Discriminate against interfering jammers through the use of multiple narrow beam antennas or adaptive antennas

4) Measure attitude

The antenna beams can be steered electronically or mechanically, if necessary. The most common GPS receivers employ only a single omnidirectional (really hemispherical antenna). The output of the antenna is then fed to a radio frequency filter/low-noise amplifier combination in order to amplify the signal and to filter out potential high-level interfering signals in adjacent frequency bands that might either saturate the amplifier or drive it into a nonlinear region of operation. The filters must be selected with low loss and sufficient bandwidth and phase linearity to minimize the distortion of the desired C/A- or P(Y)-code (precise nonstandard code) signals. The signal then passes through serial stages of radio frequency amplification, downconversion, and intermediate frequency (IF) amplification and sampling/quantizing. The sampling and quantizing of the signal can either be performed at IF or at baseband. In either case, in-phase and quadrature (I, Q) samples are taken of the received signals plus noise. At the present state-of-the-art, the functions of radio frequency amplification, downconversion, IF amplification, and A/D sampling can be implemented with a single MMIC (monolithic microwave integrated circuit) chip.

The I, Q samples are then fed to a parallel set of DLLs each of which tracks a different satellite signal and recovers the carrier, which is bi-phase modulated with both the GPS codes and the GPS navigation data. The DLL[2-4] and associated demodulators provide estimates of the pseudorange, carrier phase, and navigation data for each satellite. Typically, the number of parallel tracking DLL varies from 2 to 16 and can possibly track all of the satellites in view at both L_1 and L_2 frequencies simultaneously (see Chapters 7 and 8, this volume). Generally, at least five satellites are tracked as a minimum, either in parallel or in time sequence. At the present state of the art, a 10-channel receiver with 10 parallel DLLs can be implemented on one CMOS chip.

The parallel measurements of pseudoranges and carrier phase along with the navigation data for each satellite are then sent to the navigation data processor where the position of each satellite is calculated from the navigation data in subframes 2,3 at the time of each pseudorange measurement (see Chapters 3 and 4, this volume, for detailed discussion of the GPS signal and navigation data). The pseudorange and phase data are then corrected for the various perturbations, including satellite clock errors, Earth rotation, ionosphere delay, troposphere delay, relativistic effects, and equipment delays. The corrected pseudorange data, phase or accumulated phase [accumulated delta range ADR] measurements along with other sensor data are then processed by the Kalman filter, which estimates user position and velocity state vector. As discussed in Chapter 9, this volume, it is also possible to integrate the Kalman filter with the DLL instead of performing these operations independently. The output of the Kalman filter estimator provides position, velocity, and time estimates relative to the user antenna phase center. These coordinates are usually computed in ECEF coordinates and are then transferred by appropriate geodetic transformation to a local coordinate set convenient to the user.

The Kalman filter may also receive inputs from various other sensors; e.g., barometric altimeters, dead-reckoning estimates of attitude, heading, speed, inertial navigation ring laser gyros, or other navigation aids. This Kalman filter estimate of user position can also be used in a differential mode or kinematic survey mode with other GPS units where at least one unit is at a known reference point, in order to provide precision geodetic survey, more accurate airborne or shipborne navigation, or a common view mode, precision differential time transfer (see Chapter 9, this volume and Chapter 1 of the companion volume). The position, velocity, and time information can then be used with other user-provided information to provide tracks of user positions vs. time, display position on a map, to show way-points to a desired destination, or to satisfy a wide variety of other applications.

B. Uses of GPS

A partial listing of the uses of GPS includes the following:

1) Aircraft navigation—GPS and differential GPS, commercial and general aviation aircraft

2) Land mobile navigation—automobiles, trucks, and buses

3) Marine vessel navigation—GPS and differential GPS

4) Time transfer between clocks

5) Spacecraft orbit determination

6) Attitude determination using multiple antennas

7) Kinematic survey

8) Ionospheric measurement

See the companion volume for in-depth discussions of many of these applications.

1. Various Applications of GPS Receivers

a. Airborne GPS receivers. Airborne GPS receivers provide three-dimensional real-time navigation. These receivers must receive GPS signals from a minimum of four satellites to solve for four unknowns (x,y,z,T) because the airborne receiver clock generally is imprecise. A fifth satellite is needed for satellite handover because periodically new satellites are coming into view while another satellite is going out of view. In addition, the receivers also generally provide three-dimensional velocity estimates and clock drift estimates by making Doppler measurements on the carrier. The GPS receiver may process these satellite signals either in parallel or in time sequence. Aircraft in banked turns may suffer blockage of one or more satellites because of wing or other obstructions. Generally, receivers operate by tracking many more than the minimum four or five satellites in parallel; e.g., 8 to 12, and the clock stability may be sufficient to flywheel through momentary satellite blockage. The GPS receiver may be augmented by inertial navigation systems and/or other navigation aids to provide a hybrid GPS/inertial solution. For higher accuracy, the GPS receiver may operate in a differential mode wherein a GPS receiver at an appropriate known site; e.g., an airport, transmits differential corrections for GPS errors. In the Wide Area

Augmentation System (WAAS), corrections are transmitted to users in the form of satellite clock and position corrections, and ionospheric delay estimates which are valid over a wider geographic region. Further applications are covered in Chapters 12–15 of the companion volume.

b. Land mobile navigation. Land vehicles may require either two- or three-dimensional positioning. Generally, the altitude is varying, but typically, its variation is at a much slower relative rate compared to its horizontal motion in contrast to an aircraft. In principle, land vehicles can operate at least for a time with only two or three satellites because of the slowly varying vertical component. A fourth satellite could provide less frequent periodic updates of the altitude. However, land vehicles also are subject to blockage or shadowing of one or more satellites by trees, hills, or man-made obstructions. Another augmentation is the use of a magnetic or gyroscopic heading indicator and a wheel counter (see Chapter 10 of the companion volume).

c. Marine navigation. Navigation on the ocean or large bodies of water is usually at an altitude that varies only with the tides and any roll, pitch, and yaw-induced motion of the GPS antenna aboard the ship. For most purposes, only two dimensions are unknown. Thus, three satellites are adequate to solve for position and two satellites can suffice if a third is employed for periodic updates of the ship receiver clock. Again, however, typical receivers may operate on all satellites in view, and differential navigation can be employed for greater accuracy in harbors. This application is discussed in Chapter 11 of the companion volume.

d. Time transfer using GPS. Time can be transferred from a reference station to a clock of known location with a single satellite. Four or more satellites can be used initially if the exact location of the remote clock is unknown initially. Greater accuracy can be obtained with the civil signal using "common view" time transfer wherein two clocks at different locations are both within line of sight of the same GPS satellite. In this instance, receivers at each site are tracking the same GPS satellite simultaneously. Satellite clock errors along with any clock dither caused by SA cancel in this mode. Some fraction of the satellite position and ionospheric errors also cancel, depending on the relative separation between the two locations. This cancellation has a residual that depends (approximately) on the size of the angle between the two locations (see Chapter 17 of the companion volume).

e. Spacecraft GPS receivers. The position/orbit of a near-Earth satellite can be determined by placing a GPS receiver onboard. If the user satellite is below the altitude of GPS, then the satellite can receive the GPS signals from satellites in view above and to the sides where not shadowed by the Earth. If the user satellite is above the altitude of GPS, for example, at geostationary orbit, the user satellite can still receive the GPS signal as it passes on either side of the Earth's shadow. The GPS satellite signals are transmitted to Earth by an antenna pattern slightly greater than the Earth angle, and therefore they extend beyond the Earth's limb. The range to the GPS satellite is, of course, approximately equal to the sum of the GPS altitude and the synchronous satellite altitude, and the GPS signal is accordingly weaker. Furthermore, the GDOPs usually will not be as good as for an Earth-based user. The weaker signal can be compensated by the gain of a directional antenna (at some cost). This approach is discussed in Chapter 21 of the companion volume.

f. Differential GPS (DGPS). If two GPS receivers operate in relatively close proximity, many of the errors inherent in two GPS position solutions are common to both solutions. For example, satellite clock time errors and a significant fraction of the satellite ephemeris, ionospheric, and other errors cancel when we seek a differential or relative position solution. Thus, if one receiver is at a known, fixed position it can transmit pseudorange correction information to other receivers in the area so they can achieve higher relative position accuracy. In the future, it is expected that many GPS receivers will operate in the differential mode as differential corrections become more available. A simplified version of DGPS operation is shown in Fig. 10. The DGPS reference station transmits pseudorange correction information ($\approx$250 bps) for each satellite in view on a separate radio frequency carrier. Because there may be a number of DGPS stations in a network, the data would typically include an almanac giving the location of other DGPS reference stations so the user can use the closest station. Differential GPS is discussed in Chapter 1 of the companion volume.

Differential GPS normally is limited to separations between users and reference stations to approximately 100 km. To carry out a similar differential GPS operation over a wider region a concept known as wide area differential GPS (WADGPS) or wide area augmentation system (WAAS) has evolved. Wide area differential GPS employs a set of monitor stations spread out geographically and a central control or monitor station in somewhat the same mold as the GPS control segment, but simpler. The WADGPS upload station then would relay GPS satellite position and clock and atmosphere corrections via separate geostationary relay satellites.

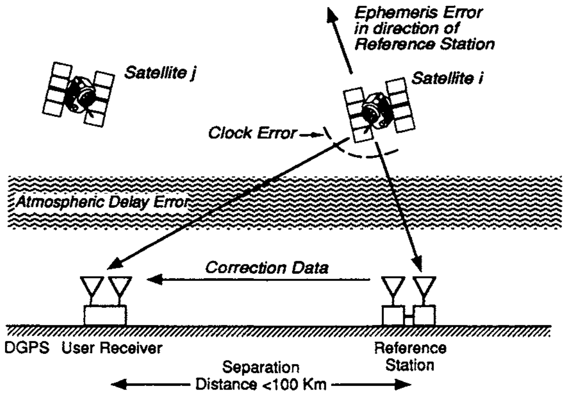


Fig. 10 Simplified view of differential GPS. This correction can completely eliminate satellite clock error offsets but ephemeris and atmospheric corrections differ for the user from the reference station by an amount that depends on the separation distance.

The corrections would be in real time with delays of less than 30 s. Wide area DGPS is discussed in Chapters 3 and 4 of the companion volume.

g. GPS survey. Global positioning system survey operates with double differencing operations similar to those described above, but carrier measurements are employed to get resolutions on the order of a fraction of the carrier wavelength. For example, the GPS carrier wavelength at 1.57542 GHz is approximately 19 cm. If a carrier phase cycle can be measured to 1%, the differential range error is only 2 mm. Certainly there are ambiguities in the carrier phase measurements, and these must be resolved with various double-differencing techniques.

The basic concept of double differencing is illustrated in simplified form is Fig. 11 where carrier phase measurements are differenced for a single satellite and the double differenced for two (or more) satellites. Obviously, this difference can be carried out at multiple time intervals to resolve the ambiguity. Survey is discussed in detail in Chapters 18 and 20 in the companion volume.

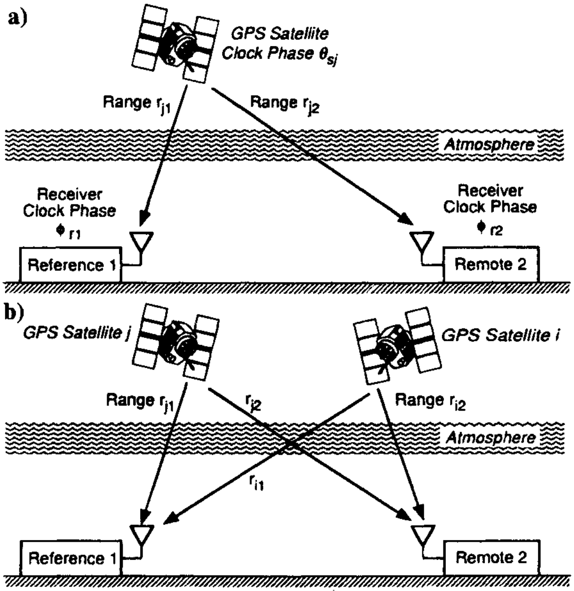


Fig. 11 Single and double carrier phase differencing for GPS and GPS-kinematic survey: a) *Single-difference receiver.* The phase received in receiver 1 is the satellite phase θ_{sj} minus the range delay effect $\omega_o r_{j1}/c$ minus the receiver reference clock phase ϕ_{r1}; namely, $\phi_{1j}(t) = \theta_{sj} - \omega_o r_{j1}/c - \phi_{r1}$. The first difference is then $\phi_{2j}(t) - \phi_{1j} = \Delta_j(t) = \omega_o(r_{j1} - r_{j2})/c + \phi_{r1} - \phi_{r2}$ where the first difference cancels out the satellite clock phase. Also, most of the atmospheric effects cancel if the separation distance is sufficiently small; b) *double-difference receiver.* The second difference is the difference between two first differences for satellite i and j; namely, $\delta_{ij}(t) = \Delta_j(t) - \Delta_i(t) = \omega_o[(r_{j1} - r_{j2}) - (r_{i1} - r_{i2})]/c$ where both the receiver carrier phase offsets cancel.

h. Attitude Determination. In addition to position measurement, GPS can also be employed in an interferometric mode with multiple user antennas to determine vehicle attitude orientation. The GPS codes can isolate each satellite signal, which can then be employed in a differential phase measurement mode with two antennas to measure angular offset for the axis of the rotation of each antenna pair and the direction of the satellite signal. Usually this technique employs carrier phase measurements. Because the baselines are short, simplified techniques to resolve the $n\lambda$ uncertainty can be used. Attitude determination is further explored in Chapter 19 of the companion volume.

i. Hybrid GPS receivers. In many applications, it makes sense to integrate GPS receiver measurements with inertial measurement systems or other navigation aids. The two sets of measurements can provide better capability than either could alone. For example, temporary shadowing or other interruption of the GPS satellite signals by a momentary obstruction or interference can be accommodated by an inertial system that can allow the navigation systems to continue to operate without interruption. (See Chapters 6–9 of the companion volume.)

VI. GPS Signal Perturbations—Atmospheric/Ionospheric/Tropospheric Multipath Effects

The GPS signal frequencies L_1 and L_2 at 1.57542 GHz and 1.2276 GHz are sufficiently high to keep the ionospheric delay effects relatively small, yet not so high as to cause too large a path loss with the use of small omnidirectional antennas (which do not require pointing). In addition, the signal frequency is not so high as to cause any significant path loss attributable to rainfall attenuation. Nonetheless, the atmosphere does cause small but nonnegligible effects. As the GPS signal passes through the atmosphere from the satellite to the user, the signal encounters a number of propagation effects, the magnitude of which depends on the elevation angle of the signal path and the atmospheric environment where the user is located. These effects include the following:

1) Ionospheric group delay and scintillation

2) Group delay caused by wet and dry atmosphere—the troposphere and stratosphere

3) Atmospheric attenuation in the troposphere and stratosphere

There are also effects caused by multipath signals from reflective surfaces and scattering. These effects are discussed fully in Chapters 12, 13, and 14, this volume; the discussion below simply introduces the key principles.

A. Ionospheric Effects

The ionosphere is a region of ionized gases that varies widely from day to day and with solar conditions and also has a large diurnal fluctuation. The presence of the ionosphere changes the velocity of propagation v according to the refractive index $n = c/v$. The cumulative effect also depends on the angle of penetration through the ionosphere as shown in Fig. 12. The refractive index $n(r)$, in turn, varies along the propagation path r. The lower extent of the ionosphere is above 75–100 km, and the ionosphere peak electron content is somewhere in

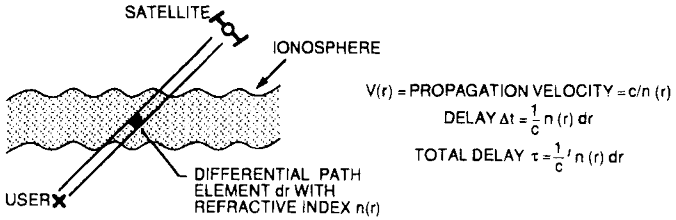

SATELLITE
IONOSPHERE
V(r) = PROPAGATION VELOCITY = c/n (r)
DELAY Δt = 1/c n (r) dr
TOTAL DELAY τ = 1/c ∫ n (r) dr
DIFFERENTIAL PATH ELEMENT dr WITH REFRACTIVE INDEX n(r)
USER

Fig. 12 Ionospheric delay along path through the medium.

the vicinity of 200–400 km. The peak ionospheric electron content can vary by as much as two orders of magnitude between day and night. A fundamental difference between the refractive index for the ionosphere and that for the troposphere is that the refractive index for the ionosphere varies with frequency because of the ionized gases.

The ionosphere can cause two primary effects on the GPS signal. The first is a combination of *group delay* and *carrier phase advance,* which varies with the exact paths and the electron density through which the satellite to user signal traverses the ionosphere. The second effect is ionospheric *scintillation,* which can at some latitudes cause the received signal amplitude and phase to fluctuate rapidly with time. Both effects depend the on the radio frequency and influence the GPS signal design. There are other effects, Faraday rotation and ray bending changing the angle of arrival, but these effects are not significant for purposes here.

At GPS frequencies in the 1.6 GHz frequency region, the ionospheric zenith path delay tends to vary with time in a diurnal fashion, as shown in Fig. 13 and might vary from 2 to 50 ns. As can be seen, however, the diurnal variation fluctuates quite markedly from day to day in a manner that seems difficult to predict. To first order, the ionospheric delay $\Delta\tau$ varies inversely with frequency squared $\Delta\tau \cong A/f^2$. As shown later, this relationship permits us to make dual frequency measurements at L_1 and L_2 to estimate the ionospheric delay. Note that because the lower extent of the ionosphere is typically well above the Earth's surface (see Fig. 14), the angles ϕ of entrance and exit of a satellite observed at 0 deg elevation angle by a user on the ground are well above 0 deg.

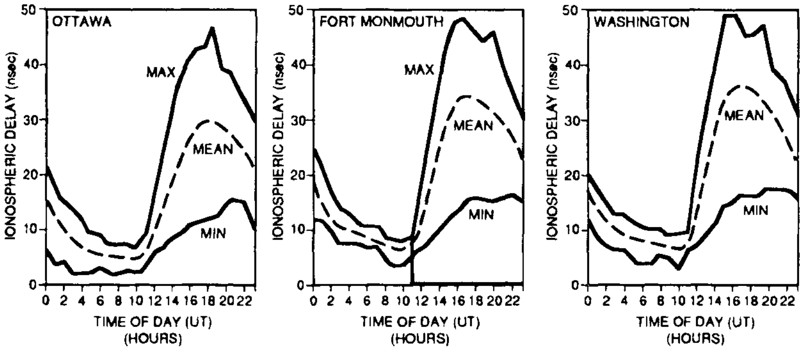


Fig. 13 Typical mean ionospheric delay and envelope of delay variation vs time of day during March 1958. Satellite at zenith f = 1.6 GHz.[5]

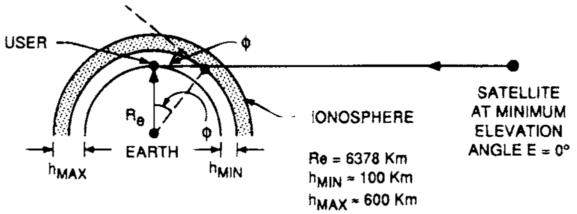


Fig. 14 Angle of incidence and exit of the ionosphere ϕ.

Thus, at lower elevation angles, the signal path transits through a larger extent of the ionosphere as shown in Fig 14. The delay at any elevation angle can be described as the ratio of actual delay to the vertical delay at $E = 90$ deg. This ratio is defined as the obliquity factor Q as shown in Fig. 15.

Note that the obliquity factor can be as high as three for low elevation angles. Because the ionosphere extends over moderate altitudes, 0.1–0.3 Earth radii, the satellite user signal path does not penetrate the ionosphere at very low elevation angles. This ionospheric effect is unlike the troposphere effects where the troposphere extends down to the Earth's surface. For an upper limit of the ionosphere of 600 km, the minimum angle of entrance to the ionosphere from the satellite is 24 deg. If the lower limit of the ionosphere is at $h = 100$ km, then the angle of exit is 10 deg. Height differences of 160–220 NM (160 NM ≅ 296 km) do not cause the obliquity factor to vary greatly at low elevation angles, as shown in Fig. 15. If the zenith ionospheric delay is 50 ns, then for an obliquity factor of three, the ionospheric delay at the lower elevation angles can be as much as 150 ns or approximately 45 m at 1.6 GHz. Clearly, this amount of unknown excess propagation delay is not consistent with a 15-m position accuracy objective for GPS and must be compensated for in some manner, either by modeling, measurement, or operation in a differential mode. Intuitively common ranging errors tend to affect the clock much more than horizontal position.

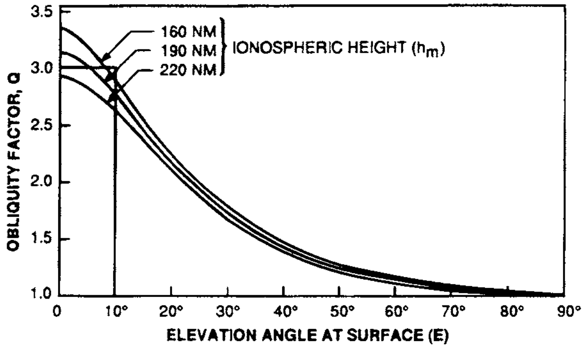


Fig. 15 Ionospheric group delay obliquity factor as a function of the relationship surface elevation angle.[5] The obliquity factor Q is the ratio of the ionospheric path delay for a satellite at elevation angle E to the delay function satellite at zenith (note 1 NM = 1.852 km).

B. Tropospheric Effects

The troposphere/stratosphere can produce a variety of propagation effects on radio waves from the satellite including the following:

1) Atmospheric attenuation
2) Tropospheric scintillation
3) Tropospheric refraction caused by the wet and dry atmosphere that produce excess delay in the signal

1. Tropospheric Group Delay

The troposphere is a region of dry gases and water vapor that extends up to approximately 50 km. This region has an index of refraction, $n(h) = 1 + N(h) \times 10^{-6}$, that varies with altitude. The index of refraction is slightly greater than unity, and hence, causes an excess group delay in the signal waveform beyond that of free space. This region is not ionized and is not frequency dispersive because the excess group delay $\Delta\tau$ is constant with frequency for frequencies below 15 GHz and is approximately equal to the following:

$$\Delta\,\tau = \int_{\text{path}} N(h)dh \times 10^{-6}$$

The excess group delay is normally on the order of 2.6 m for a satellite at zenith, and it can exceed 20 m at elevation angles below 10 deg. Thus, it must be modeled and removed if high accuracy positioning and time transfer are to be achieved. Dry atmospheric effects that are relatively easily modeled account for approximately 90% of the tropospheric excess delay. The wet atmosphere, although only about 10% of the total, is highly variable and very difficult to model. Detailed discussions of the ionospheric and tropospheric effects on GPS are given in Chapters 12 and 13, this volume.

C. Multipath Effects

1. Multipath

Some of the most difficult navigation problems are for aircraft. Aircraft navigation and three-dimensional navigation in general are also prime motivations for the GPS system. It is important to examine potential multipath effects that can be present in aircraft navigation. An aircraft flying at altitude h has a multipath environment with ground or sea surface reflections, as shown in Fig. 16. If the satellite elevation angle is E, the reflected ray is delayed with respect to the direct ray by $\Delta R = c\Delta\tau = 2h \sin E$. If, as an example, $h = 1$ km and $E = 10$ deg, then the delay difference in the reflected signal is $\Delta\tau = 1.16$ μs. The sea surface reflection cannot be avoided easily by antenna design if we must operate with satellites at low or moderate elevation angles. (The antenna pattern must allow for aircraft banking.) Furthermore, the reflected signal amplitude from the sea surface can at times be nearly as large or sometimes even larger than the direct ray. As shown later in Chapters 3 and 7, the GPS receiver can effectively reject most of the multipath signal if the differential delay $\Delta\tau > 1.5$ μs for the C/A code and 0.15 μs for the P(Y)-code. Note the region of potential

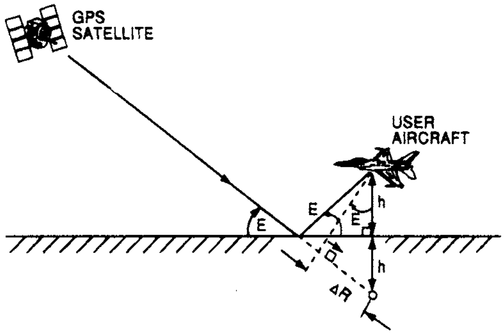


Fig. 16 Multipath effects. Multipath delay varies with elevation angle *E* and user altitude *h*. Delay = $\Delta R = 2h \sin E$.

mutipath delay problems for the C/A code is then

$$1.5\ \mu\text{s} > \frac{2b}{c} \sin E = \Delta\tau$$

or

$$h \sin E < (1.5\ \mu\text{s})c = 448.5\ \text{m}.$$

Because the satellite is moving, the multipath will, in general, be time varying. A reflected multipath signal has the following Doppler shift

$$\Delta f = \Delta\dot{\phi} = \frac{2\dot{h} \sin E + 2h\, \dot{E} \cos E}{\lambda}$$

for the example shown in Fig. 16. More generally, the multipath may consist of a whole array of scatterers/reflectors, as shown in the impulse response of Fig. 17b. If the number of reflectors is sufficiently large and they are modeled as

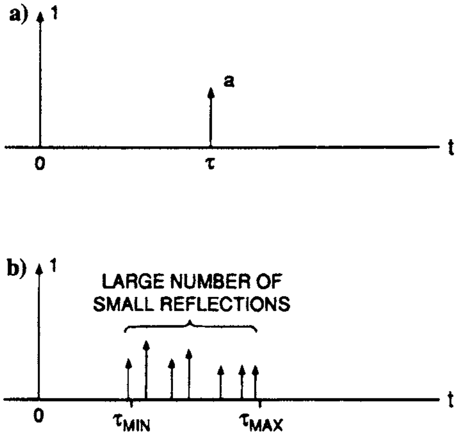


Fig. 17 Multipath channel impulse response with a) a single specular reflection; and b) a large number of smaller reflections.

independent random reflections, then the summed multipath reflection is approximately Gaussian and has a Rayleigh distribution in amplitude.

Within the constraints of available bandwidth (≈2 MHz per channel for the usual civilian receivers) and limitations on receiver complexity, the GPS signal is designed to resist the interference from multipath signals for delay differences that exceed 1 μs. Of special importance are multipath signals with a delay difference corresponding to the aircraft sea surface reflections described above. The capability for multipath discrimination of the selected GPS signal is examined in Chapter 14, this volume. We should recognize however, that there can also be reflections from nearby metallic or conducting surfaces (e.g., aircraft wings or stabilizers) that cannot be discriminated against by choice of signal within the bandwidth constraint because the delay difference is too small.

D. Other Perturbing Effects

In addition to the satellite clock and ephemeris errors of the GPS control segment and the atmospheric effects of the ionosphere, troposphere, and multipath, there are several other effects that are important for at least some users. Each of these is briefly discussed below.

1. Relativistic Effects

The Global Positioning System is perhaps the first widely used system where relativistic effects are not negligible. For example, referring back to Figs. 2 and 4, the position and time were all discussed in gravity-free inertial space with stationary users and a nonrotating Earth. In fact, there are several relativistic effects that are nonnegligible. These effects include the following: 1) gravitational field of the Earth and Earth rotation; and 2) velocities of satellite and user. The major effects cause an average increase in the satellite clock frequency as observed by a user on Earth. These effects are partially accommodated by purposely setting the satellite clock slightly low in frequency prior to launch. This topic is discussed in Chapter 18, this volume.

2. Foliage Attenuation

One of the major classes of users for GPS are ground mobile users. If these users are traveling along a road or highway, there is the possibility of obstructions or tree foliage attenuations of the GPS segments. This topic is discussed in Chapter 15, this volume.

3. Selective Availability

To reduce the potential for GPS to be used in hostilities toward the United States, the accuracy of the GPS signal for civil users can be purposely reduced by a capability called selective availability, which is discussed in Chapter 17, this volume.

References

[1]Kaplan, G. H., "The IAU Resolutions on Astronomical Constants, Time Scales, and the Fundamental Reference Frame," U.S. Naval Observatory Circular 163, Washington, DC, Dec. 1981.

[2] Spilker, J. J., Jr., and Magill, D. T., "The Delay Lock Loop—An Optimum Tracking Device," *Proceedings of the IRE,* Vol. 49, Sept. 1961.

[3] Spilker, J. J., Jr., *Digital Communications by Satellite,* Prentice-Hall, Englewood Cliffs, NJ, 1977, 1995.

[4] Spilker, J. J., Jr., "GPS Signal Structure and Performance Characteristics," *Navigation,* Summer 1978, pp. 121–146; also published in *Global Positioning System Papers,* Institute of Navigation, Washington, DC.

[5] National Research Council, *The Global Positioning System—A Shared National Asset,* National Academy Press, Washington, DC, 1995.

Chapter 3

GPS Signal Structure and Theoretical Performance

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

THIS chapter discusses the details of the Global Positioning System (GPS) signal structure, its specifications, and general properties. The chapter begins with a review of the general performance objectives and quantitative requirements of the signal. Because the GPS signal falls into a broad category of signals known as spread spectrum signaling, the fundamentals of spread spectrum signaling are introduced. The chapter continues with a detailed description of the GPS signal structure for both the precision (P code) and civil coarse/acquisition (C/A code) signals. The various minor distortions and imperfections permitted by the GPS satellite–user interface specification are also discussed. Although the general format of the navigation data is summarized, details of the navigation message are given in the next chapter. The radio frequency signal levels and relevant signal-to-noise ratios are discussed next. The chapter concludes with a discussion of the signal performance characteristics including: 1) C/A- and P-code cross-correlation properties for multiple access, and 2) performance bounds on the C/A- and P-code pseudorange tracking accuracy for the optimal delay–lock-loop tracking receivers.

A brief summary of Galois fields that are the mathematical basis for maximal length and Gold sequences is given in the Appendix.

A. Summary of Desired GPS Navigation Signal Properties

Based on the navigation accuracy and system requirements and the relevant physics/communication theory discussed in the previous chapter, the system level accuracy requirements can now be stated and translated into signal measurement accuracy requirements. User position and velocity accuracy objectives translate into requirements on pseudorange and other GPS signal measurements and infor-

*Ph.D., Chairman of the Board.

mation on satellite position and clock time at the time of transmission that must be available to the user receiver. This required information is summarized in Fig. 1.

The accuracy of the pseudorange measurements can be related to the desired accuracy of position by the various **dilutions of precision** (DOP); e.g., position dilution of precision (PDOP), and horizontal dilution of precision (HDOP). If it is assumed that PDOP ≈ 3 then a 10-m rms user position error translates into a pseudorange accuracy* of ≤ 10 m/3 $= 3.33$ m or 11 ns. Likewise, a civil user needing 100-m real-time accuracy translates to a pseudorange accuracy of roughly 110 ns. In addition, the GPS signal should also possess the following properties:

1) Tolerance to signals from other GPS satellites sharing the same frequency band; i.e., multiple access capability

2) Tolerance to some level of multipath interference—there are many potential sources of multipath reflection; e.g., man-made or natural objects or the sea surface for an aircraft flying over water

3) Tolerance to reasonable levels of unintentional or intentional interference, jamming, or spoofing by a signal designed to mimic a GPS signal

4) Ability to provide ionosphere delay measurements—dual frequency measurements made at L_1, L_2 frequencies must permit accurate estimation of the slowly changing ionosphere

1. *Flux Density Constraints*

In addition to the requirements stated above for the GPS signal, there are requirements that the GPS signal received on the Earth be sufficiently low in power spectral density so as to avoid interference with terrestrial microwave line-of-sight communication. For example, a line-of-sight microwave terminal carrying a large number of 4-kHz voice channels potentially can receive interference from a GPS satellite signal that might be observed within the antenna pattern of the receive microwave antenna. Thus, the power flux density of the GPS

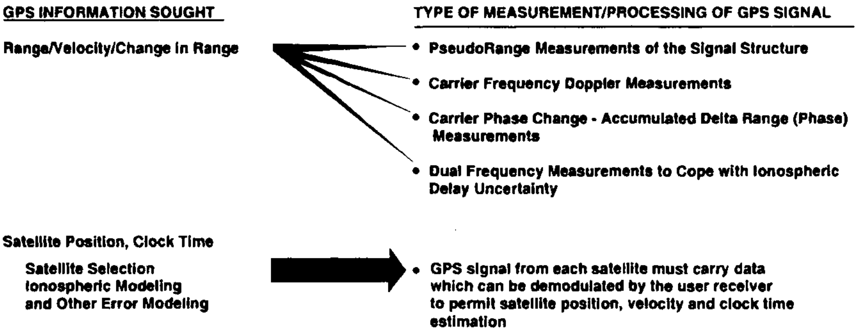

GPS INFORMATION SOUGHT
TYPE OF MEASUREMENT/PROCESSING OF GPS SIGNAL
Range/Velocity/Change in Range
• PseudoRange Measurements of the Signal Structure
• Carrier Frequency Doppler Measurements
• Carrier Phase Change - Accumulated Delta Range (Phase) Measurements
• Dual Frequency Measurements to Cope with Ionospheric Delay Uncertainty
Satellite Position, Clock Time
Satellite Selection
Ionospheric Modeling
and Other Error Modeling
• GPS signal from each satellite must carry data which can be demodulated by the user receiver to permit satellite position, velocity and clock time estimation

Fig. 1 GPS information sought and measurement/signal processing by the user receiver.

*This statement assumes that there are no other error effects besides pseudorange. There are other errors; e.g., satellite position and clock time errors, but at the moment only the GPS pseudorange error effects are considered.

satellite signal in a 4-kHz band is constrained so as to remain below a certain level, thus eliminating the possibility of interference on one or more voice channels. Because the 24 GPS satellites orbit the Earth in 12-h orbits, many microwave radio locations will observe GPS satellites at low-elevation angles at one time or another. For this reason, the International Telecommunication Union (ITU) has set flux density regulations on the power that can be generated by a satellite-to-Earth link. In the 1.525–2.500 GHz band, the flux density limit for low-elevation angles is -154 dBW/m^2 for any 4/kHz frequency band[1]. Because the constraint is on power flux spectral density rather than total radiated power, a satellite can radiate more total power and stay within the flux density limit if the signal energy is spread out fairly uniformly over a wider spectral band. The Global Positioning System uses spread spectrum signals to achieve this goal wherein the signal spectra are spread out over a much wider bandwidth than their information content in order to permit use of higher power levels and, of course, to achieve sufficiently precise ranging accuracy. For a unity (0 dB) gain antenna, the aperture area is $A = \lambda^2/4\pi$, and for GPS L_1 at 1.57542 GHz, where the wavelength is $\lambda = 0.1904$ m and $A = 2.886 \times 10^{-3}$ m^2 or -25.4 dB relative to 1 m^2. Thus, this flux density limit transblates to a power level to a unit gain antenna at 1.54542 GHz of $-154 - 25.4 = -179.4$ dBW in any 4-kHz frequency band.

In addition to constraints for line-of-sight microwave radio, there are also constraints to protect radio astronomy. Radio astronomy makes use of the 1420.4 MHz spectral line of neutral atomic hydrogen (the 1400–1427 MHz band is assigned for radio astronomy) and the OH radical molecule with lines at 1612.232, 1665.402, 1667.359, 1720.530 MHz.[2] Thus, the GPS satellite signal is specially filtered to avoid interference with these bands.

B. Fundamentals of Spread Spectrum Signaling

Spread spectrum signaling in its most fundamental form is a method of taking a data signal $D(t)$ of bandwidth B_d that is modulated on a sinusoidal carrier to form $d(t)$, and then spreading its bandwidth to a much larger value B_s where $B_s >> B_d$. The bandwidth spreading can be accomplished by multiplying the data-modulated carrier by a wide bandwidth-spreading waveform $s(t)$. A simplified spread spectrum system is shown in Fig. 2. The figure shows a conventional biphase modulated transmitter* on the far left, followed by a spectrum-spreading operation, an additive noise and interference transmission channel, and the spread spectrum receiver processor. A binary data bit stream $D(t)$ with values $D = \pm 1$ and clock rate f_d is first modulated on a carrier of power P_d to form the narrow bandwidth signal:

$$d(t) = D(t)\sqrt{2P_d} \cos \omega_o t \tag{1}$$

This narrow bandwidth signal of bandwidth B_d is then spread in bandwidth by a binary pseudorandom signal $s(t)$ where $s(t) = \pm 1$ and has a clock rate f_c that greatly exceeds the data bit rate; i.e., $f_c >> f_d$. For random data and spreading

*In general the data signal $D(t)$ can be multilevel and complex. However, this discussion is restricted to binary real $D(t) = \pm 1$.

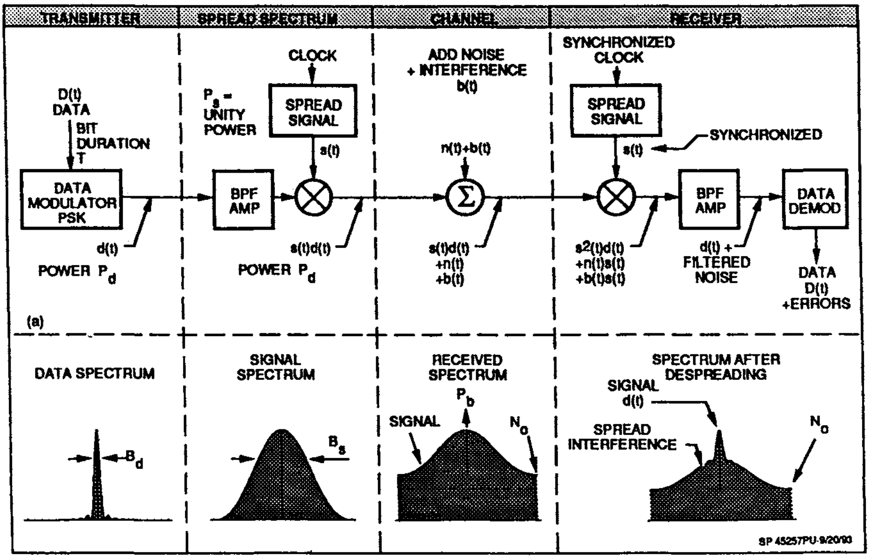


Fig. 2 Simplified spread spectrum link.

codes, the data $D(t)$ and spreading waveforms $s(t)$ have the following power spectral densities, respectively, as shown in Fig. 2.

$$G_d(f) = \frac{1}{f_d}[(\sin \pi f/f_d)/(\pi f_d)]^2 \quad \text{and} \quad G_s(f) = \frac{1}{f_c}[(\sin \pi f/f_c)/(\pi f/f_c)]^2$$

Because the timing of the data and clock transitions are synchronous, the spread spectrum product $D(t)$ $s(t)$ has exactly the same spectrum as that of $s(t)$ alone. The spread spectrum signal then has the following form:

$$s_o(t) = s(t)d(t) = s(t)D(t)\sqrt{2P_d}\cos \omega_o t \tag{2}$$

where the spreading signal is as follows:

$$s(t) = \sum_{n=-\infty}^{\infty} S_n p(t - nT_c) \tag{3}$$

and $p(t)$ is a rectangular unit pulse over the interval $\{0, T_c = 1/f_c\}$ and S_n is a random or pseudorandom sequence $S_n = \pm 1$. In general, $p(t)$ can represent a filtered pulse, and different spreading waveforms $s_i(t)$ with coefficients S_{in} in Eq. (3) can separately modulate in-phase and quadrature carrier components. For this example, we restrict the discussion to rectangular pulses and biphase modulation.

This form of spread spectrum is termed **direct sequence-spread spectrum** (DS-SS), and it is one of many different forms of spread spectrum. Other forms* include spreading by pseudorandom frequency hopping, termed frequency hop-

*There are other means for bandwidth spreading, such as low rate error correction coding, that do not employ an independent spreading waveform.[3]

spread spectrum (FH-SS), and various hybrid forms of DS-SS and FH-SS. We restrict our discussion here to DS-SS because it provides a means to recover precise timing, and at the same time, it permits recovery of the pure rf carrier. The ability to recover pure carrier is key to precision differential delay and Doppler measurements that provide accuracies on the order of 1% of a carrier wavelength.

The DS-SS signal in Fig. 2 next passes through a channel (with zero channel delay for simplicity and without loss of generality) with additive white noise $n(t)$ of power spectral density N_0 and interference $b(t)$ to form the received signal $r(t) = s(t)\, d(t) + n(t) + b(t)$ where $b(t)$ is a pure tone interference of power P_b. In the receiver, an identical and precisely time-synchronized replica spreading signal $s(t)$ is generated and correlated (multiply and filter) with the received noisy signal. The fact that the received replica must be accurately time synchronized is shown later to be the exact property that enables the system to extract accurate timing and ranging information. That is, the signal has a narrow autocorrelation envelope of width inversely proportional to the clock rate f_c. The receiver multiplier converts the desired signal $d(t)\, s(t)$ to $d(t)\, s^2(t) = d(t)$ because $s^2(t) = 1$; that is, it compresses the spread spectrum signal to its original narrow bandwidth with only the data modulation remaining. The noise spectral density is still N_0 because convolving white thermal noise with a continuous constant envelope spread spectrum signal is still white Gaussian noise. The narrowband interference $b(t)$ has now been spread to look like $s(t)$ and has bandwidth B_s, similar to the manner in which the narrowband signal $d(t)$ was spread in the transmitter. Filtering this multiplier output through a bandpass filter passes the narrowband signal $d(t)$ relatively undistorted, however only a fraction of the noise and interference power passes through the bandpass filter with power $N_0\, B_d$ and $P_b\, (B_d/B_s)$, respectively.

Demodulation of this filtered output then produces a bit error rate that is determined by this noise and interference level. Note that if there is only thermal noise and no other interference is present, then the receiver output is exactly the same in terms of signal power and noise density as if there had been no spectral spreading at all. That is, the effects of the spreading and despreading by the binary pseudorandom code $s(t)$ in the transmitter and receiver cancel. Thus, the use of properly synchronized spread spectrum signaling neither improves nor degrades the signal performance against a thermal noise background.

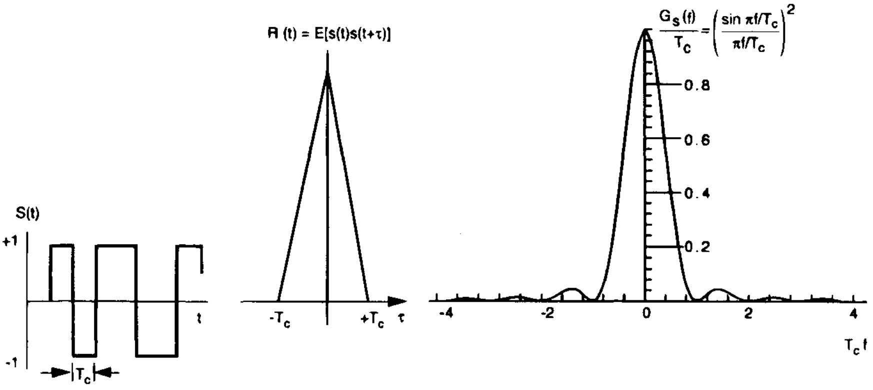
However, the performance against a tone interference of fixed power is greatly improved because the interference power level is reduced by the ratio of the clock rates f_c/f_d. The ratio, f_c/f_d, of the PN chip rate f_c to the data bit rate f_d is termed the *processing gain of* the spread spectrum system and is a key parameter because it determines what fraction (f_d/f_c) of the interference power passes through to the output. Whereas the thermal noise power increases in direct proportion to radio frequency bandwidth, the interference power is fixed and independent of bandwidth. In fact, spread spectrum signaling is effective against a much broader class of interference than simple tone interference.

1. Direct Sequence-Spread Spectrum Signals—Autocorrelation Function and Spectrum

A noise-like pseudorandom binary spreading sequence $s(t)$ bears a close resemblance to a random sequence. A purely random binary sequence is generated by a coin-flipping operation where the outcome is equally probable ±1. This sequence at clock rate f_c can generate the waveform $s(t)$ of (3), which is shown in Fig. 3a, and it has a triangular autocorrelation function and $(\sin \pi\tau f/\pi\tau f)^2$ shaped power spectral density as shown in Fig. 3b,c. As is shown in a later paragraph, a close approximation (pseudorandom) to a random sequence can be generated by using suitable feedback shift registers. Thus, we can generate a replica waveform at the receiver and suitably time synchronize this replica to the received signal.

2. *Multiple Access Performance of Spread Spectrum Signaling*

It is often desired to provide a method by which multiple signals can simultaneously access exactly the same frequency channel with minimal interference between them. Spread spectrum signaling has the capability to provide a form of multiple access signaling called code division multiple access (CDMA) wherein multiple signals can be transmitted in exactly the same frequency channel with limited interference between users, if the total number of user signals M is not too large. This multiple access capability is important for GPS because a user receiver may receive simultaneously 10 GPS signals from 10 different satellites, wherein all signals occupy the same frequency channel and are continuous (i.e., not time gated). For example, assume that there are M signals, all of exactly the same power P_s received at a receiver antenna. If all M signals are received with exactly the same code clock delay, it is possible to select a certain number of signals that are completely orthogonal, and thus cause no multiple access interference provided $M \leq f_c T_d$ where $f_d = 1/T_d$ is the data bit rate. However, in many communication/ranging tasks orthogonal signaling is not possible because the signal sources—the GPS satellites in our example—cannot possibly all be equally distant from each user. Good multiple access performance is still possible, however, by selecting the different GPS spread spectrum pseudorandom codes to be nearly uncorrelated for all possible time offsets.



a) b) c)

Fig. 3 Random binary sequence, autocorrelation function, and power spectral density—a) waveform, b) autocorrelation function, and c) power spectral density. The clock rate is $f_c = 1/T_c$.

Examine two multiple access signals $s_i(t)$ and $s_j(t)$, which are both uncorrelated pseudorandom codes with identical spectra $G_s(f)$ and are both transmitted on the same frequency channel and received with independent random timing. The receiver of Fig. 2 for the first signal; e.g., $s_i(t)$, cross-correlates the received additive signals with the desired reference signal code $s_i(t)$. Ignoring the data modulation, carriers, and noise for the moment, the correlator output is then $s_i(t - \tau_1)\,[s_i(t - \tau_1) + s_j(t - \tau_2)] = 1 + s_i(t - \tau_1)\,s_j(t - \tau_2)$. The unity term is the desired component, and the spread spectrum multiple access interference term is $s_i(t - \tau_1)\,s_j(t - \tau_2)$. For random time offsets between the two signals and power level P_s, the multiple access interference spectrum is defined as $G_{ma}(t)$ and is obtained by convolving the individual spectra (see Fig. 4):

$$G_{ma}(f) = P_s \int G_s(\nu)G_s(\nu - f)d\nu \quad \text{and} \quad G_{ma}(0) = P_s \int G_s^2(\nu)d\nu$$

Assume that the processing gain is large; i.e., $f_c/f_d >> 1$. Then only the multiple access interference spectrum near $f = 0$ is significant because the correlation filters can have a bandwidth on the order of f_d. The convolved spectrum at $f = 0$ can be computed to be:

$$G_{ma}(0) = P_s \int_0^\infty \left(\frac{\sin \pi f/f_c}{\pi f/f_c}\right)^4 \mathrm{d}f$$

$$G_{ma}(0) = \left(\frac{2}{3}\right)(P_s/f_c)$$

Note that if the multiple access signal has transitions coincident with the reference signal; that is, a nonrandom time offset, the multiple access interference

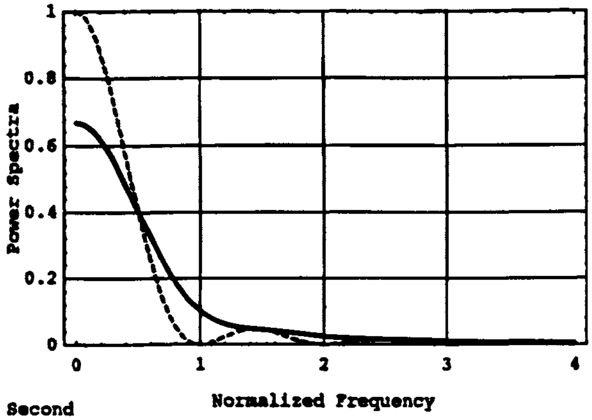


Fig. 4 Original spectrum $(\sin \pi/\tau f)^2$ (dashed line) and the convolved spectrum G_{ma} (f) (solid line) at the correlator output, for a normalized clock rate $f_c = 1$ for an unfiltered pseudorandom signal. Note that the multiple access power spectral density at $f = 0$ decreases to 2/3.

is not spread, and the factor of 2/3 does not appear. Note also, that this result (Fig. 4) assumes that the $[(\sin \pi f/f_c)/\pi f/f_c]^2$ spectrum includes all of its sidelobes and is not filtered. If the signal is filtered to include only the main lobe the factor of 2/3 increases to approximately 0.815. If the signal spectrum is rectangular the factor is unity. As shown later, the transmitted GPS C/A code is nearly unfiltered and contains sidelobes out to the 10th. The GPS P code has the same bandwidth but 10 times the clock rate, and thus only contains the mainlobe.

Because there are M-1 interfering multiple access signals, and there is only one desired signal for each tracking receiver, the net effect of the M-1 multiple access signals is to increase the effective noise spectrum in the vicinity of the desired data modulation from a value N_o with no multiple access interferences to an equivalent noise density:

$$N_{oeq} = N_o + \frac{2}{3}\left(M - 1\right)P_s/f_c = N_o[1 + \frac{2}{3}(M - 1)P_s/f_cN_o] \qquad (4)$$

Table 1 summarizes the equivalent noise density relationships for spread spectrum multiple access signals for the complete $(\sin x/x)^2$ spectrum, $(\sin x/x)^2$ mainlobe only, and rectangular spectra.

Thus, if all sidelobes are included, the effective noise density is increased by the factor $1 + (2/3)\,(M - 1)\,P_s/f_c$, and the effective energy per bit E_b to equivalent noise density ratio decreases to the following:

$$\frac{E_b}{N_{oeq}} = \frac{P_sT_d}{N_o\left[1 + \frac{2}{3}(M - 1)P_s/f_cN_o\right]} = \frac{P_s}{N_of_d}\,\frac{1}{\left[1 + \frac{2}{3}(M - 1)P_s/f_cN_o\right]} \qquad (5)$$

The quantity E_b is the energy per bit $E_b = P_s/f_d$. The quantity E_b/N_{oeq} determines the output error rate. For biphase modulated (antipodal) signaling the E_b/N_{oeq} needs to be on the order of 10 if no error correction coding is employed.* If the performance of the system is not to be degraded by more than 3 dB relative to thermal noise performance, then from (5) the number of equal power multiple access signals is limited by the following:

$$M < \frac{3}{2}\left(\frac{N_of_c}{P_s}\right) + 1 = \frac{3}{2}\left(\frac{N_of_c}{E_sf_d}\right) + 1 = \frac{3}{20}\left(\frac{f_c}{f_d}\right) + 1 \quad \text{for} \quad \frac{E_b}{N_{oeq}} = 10 \qquad (6)$$

the limit on M increases as the spread spectrum clock rate of f_c increases.† Again, it is pointed out that the factor of 3/2 applies only for the unfiltered $(\sin x/x)^2$ spectrum.[4] Note that for GPS there are often $M = 9$, 10 GPS signals in view. See the later chapter on the GPS constellation. Furthermore, note that if the

*Although GPS does not employ error correction coding, it should be pointed out that the use of spread spectrum signaling generally does permit the use of low rate error correction codes that can allow operation at low values of E_b/N_o without suffering from bandwidth expansion because the spread spectrum signaling is already broadening the spectrum by itself.

†For example, with $f_c = 10^6$ and $f_d = 50$, this result, (6), becomes:

$$M < (3/20)(10^6/50) + 1$$
$$M < 3 \times 10^3 + 1$$

Table 1 Equivalent noise density for M equal power spread spectrum signals of different power spectral densities[a]

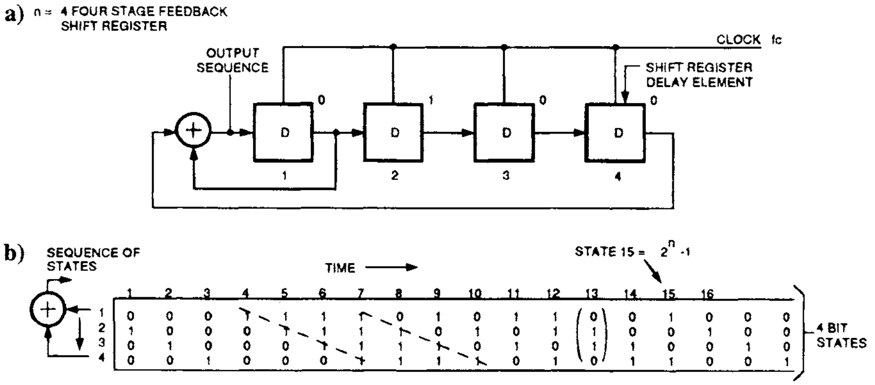
Spread spectrum signal spectra	Multiple access equivalent noise density at $F = 0$
$G_s(f)$, $-\infty < f < \infty$	$N_o\left[1 + \frac{(M-1)P_s}{N_o}\right]\int_{-\infty}^{\infty} G_s^2(f)\,\mathrm{d}f$
$\frac{1}{f_c}\left[\frac{\sin \pi f/f_c}{\pi f/f_c}\right)^2$, $-\infty < f < \infty$ All sidelobes	$N_o\left[1 + \left(\frac{2}{3}\right)\frac{(M-1)P_s}{f_c N_o}\right]$
$\frac{1}{f_c}\left[\frac{\sin \pi f/f_c}{\pi f/f_c}\right]^2\left(\frac{\pi}{2\ \text{sin integral}(2\pi)}\right)$, $-f_c < f < f_c$ Mainlobe only	$N_o\left[1 + \frac{\pi(M-1)P_s}{3f_o N_o}\cdot\left(\frac{2\ \text{sin integral}(4\pi) - \text{sin integral}(2\pi)}{(2\ \text{sin integral}(2\pi))^2}\right)\right]$ $= N_o\left[1 + \frac{(M-1)P_s}{f_o N_o}(0.815497)\right]$
Rectangular spectrum $\frac{P_s}{2f_o}$, $-f_o < f < f_o$	$N_o\left[1 + \frac{(M-1)P_s}{f_o N_o}\right]$

[a]The reference signal spectra $G_s(f)$ are all normalized to unity signal power. Each of M received signals has power P_s.

desired signal is 1/10th of the power of the other signals, the value of M permitted decreases approximately by a factor of 10. As shown later, in some instances the DS-SS code has a relatively short period (e.g., the GPS C/A code), the spectrum of the spreading code has line components spaced at the epoch rate, and the performance is not quite as good as indicated above.

3. Generation of the Spreading Signal Using Linear Feedback Shift Registers

Figure 5 shows a simple four-stage linear feedback shift register with taps after stages 1 and 4, which are modulo-2 added to form a short period maximal length pseudorandom or pseudonoise PN sequence. The sequence of shift register state vectors is shown in Fig. 5b, where the initial state vector is as follows:



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0	0	0
2	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0	0
3	0	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0
4	0	0	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1

Fig. 5 Generation of a PN sequence using a maximal length feedback shift register and the $2^4 - 1 = 15$ sequence of code generator states. The state vectors each have four binary components, as shown in b.

$$\mathbf{s}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

The state vector components are defined by the state of each of the four binary shift register delay elements. As long as this shift register is not set to the "all zero" state, it will cycle through all $2^4 - 1 = 15$ state vectors in a periodic manner. Only specific tap combinations produce a sequence of length $2^4 - 1 = 15$. In general, an n stage linear feedback shift register (LFSR) with proper taps generates all 2^n of the n-bit states except the all-zero state, thus it cycles through each of the possible state vectors. The use of the word "linear" refers to the restriction on the logic to modulo-2 adders. Thus, there are $2^n - 1$ states in the period of the sequence, and the sequence has a period $2^n - 1$.

Figure 6 shows the PN sequence at the output of a selected stage. The autocorrelation of the PN sequence where $s(t) = \pm 1$ is the following:

$$R(i) = (1/15) \int_0^{15} s(t)s(t + i)\mathrm{d}t$$

and for rectangular pulses has the form shown in Fig. 6c when expressed as a function of continuous time.

The maximal length sequence can be designed to have an arbitrarily long period $2^n - 1$ by increasing the number of stages n. As n increases, the maximal length sequence becomes more random in appearance or pseudorandom, and its spectrum approximates the $(\sin x/x)^2$ spectrum of Fig. 3. Because of the sharp (narrow in time) peak of the autocorrelation function for a high clock rate signal, the waveform also can be used for very accurate measurements of time and range or pseudorange. Obviously, this characteristic is key for GPS. An introduction

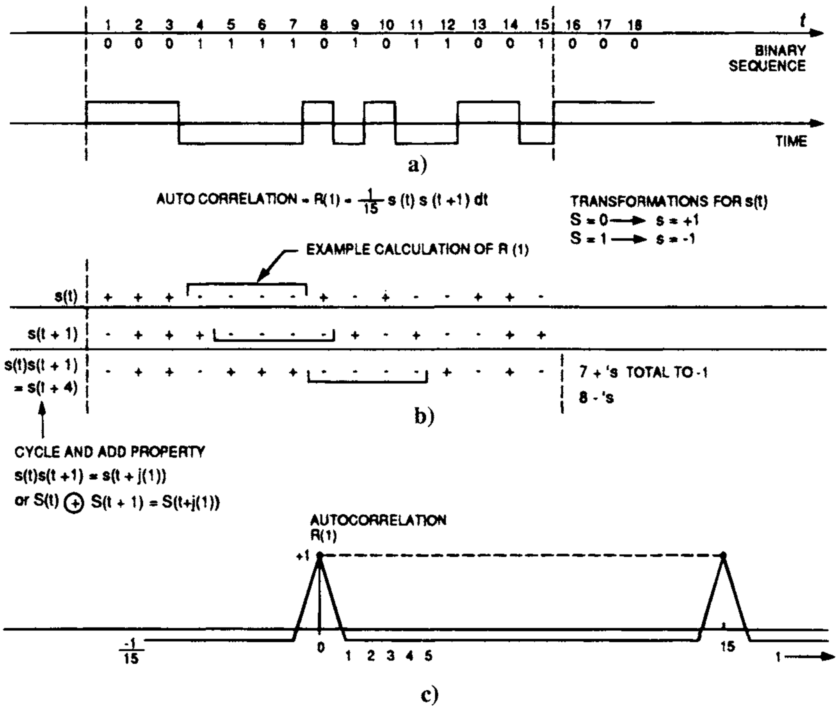

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
t
0 0 0 1 1 1 1 0 1 0 1 1 0 0 1 0 0 0
BINARY SEQUENCE
TIME
a)
AUTO CORRELATION = R(1) = 1/15 s (t) s (t +1) dt
TRANSFORMATIONS FOR s(t)
S = 0 → s = +1
S = 1 → s = -1
EXAMPLE CALCULATION OF R (1)
s(t) + + + - - - - + - + - - + + -
s(t + 1) - + + + - - - - + - + - - + +
s(t)s(t + 1) = s(t + 4) - + + - + + + - - - - + - + -
7 +'s TOTAL TO -1
8 -'s
b)
CYCLE AND ADD PROPERTY
s(t)s(t +1) = s(t + j(1))
or S(t) ⊕ S(t + 1) = S(t+j(1))
AUTOCORRELATION R(1)
+1
-1/15
0 1 2 3 4 5
15
t
c)

Fig. 6 Autocorrelation function of a PN sequence with $P = 2^n - 1 = 15$ states.

to the more detailed analysis of PN codes and Galois field algebra for PN codes is given later in the chapter as an appendix.

II. GPS Signal Structure

In this section, the structure of the GPS signal is described in detail. The general properties of the civil and precision direct sequence spread spectrum signals and codes are discussed and related to many of the system requirements. Detailed analyses of the performance of the signal are given in the next section.

As stated earlier, the GPS system must provide authorized users with a 10-m or less rms position error, which for a PDOP $\approx$ 3 translates to a required accuracy of pseudorange measurement on the order of 11 ns. The Global Positioning System chose to accomplish this required accuracy with a 10.23-Mcps precision P code. Two other GPS objectives, rapid acquisition of the P-code and providing a lesser but still revolutionary three-dimensional accuracy for the civil user are achieved by the use of the civil coarse/acquisition (C/A)-code, which has a 1.023-Mcps chip rate and a code period of 1023 chips. Civil users do not have access to the P(Y) code when the P code is in the antispoof (AS) Y-code mode. The somewhat unusual code rates of 1.023 Mcps and 10.23 Mcps are selected so that the period of the C/A code corresponds exactly to 1 ms for time-keeping purposes.

A. Multiplexing Two GPS Spread Spectrum Signals on a Single Carrier and Multiple Access of Multiple Satellite Signals

Two important questions to deal with are how to multiplex the two codes, C/A and P, on a single carrier and how to provide the multiple access of the various GPS signals that are to be received from the different satellites within the available frequency band.

The GPS L_1 signal has two spread spectrum signals, civil, C/A, and precision, P, multiplexed onto a single radio frequency carrier. In addition, the signals from multiple satellites must share the same frequency channel. The Global Positioning System multiplexes the civil and precision code on a single carrier in phase quadrature and then employs CDMA so that the different satellite signals can share the identical frequency band. Each satellite P signal occupies the entire available bandwidth to maximize timing accuracy. Table 2 shows the multiplexing and multiple access alternatives considered during the original design of the GPS signal.

Time multiplexing of the two signals, civil (coarse) and precision on one carrier; i.e., transmitting a portion or all of the period of the civil signal followed by a portion of the long military signal, was a possible choice for GPS. However, a time multiplexed signal would not have permitted continuous phase measurement of the carrier because the civil user does not have access to the military precision signal. The ability to perform precision carrier phase measurements was always considered to be of key value to the GPS system. Time gating of the shorter C/A civil signal would also change its autocorrelation characteristics and results in a less desirable cross-correlation performance.

The alternative selected for GPS is to modulate the civil C/A signal on the in-phase component of the L_1 carrier and modulate the precision P signal on a quadrature phase (90 deg rotated), thus providing a constant envelope modulated carrier even if the two signals have different power levels. The GPS signal then has the form (neglecting data modulation) $XP_i(t)\cos\omega_o t + XG_i(t)\sin\omega_o t$ where XP_i represents the P-code and XG_i represents the C/A code. Data biphase modulates both inphase quadrature components identically.

The selected multiple access technique for GPS is CDMA wherein the signals are separated through the use of codes with good cross-correlation properties. Code division multiple access in some systems has a so-called "near–far" problem when substantial differences exist in the received signal levels from different

Table 2 Alternative multiplexing and multiple access techniques considered for the global positioning system

Methods of multiplexing onto a single carrier	Time multiplex civil and military codes		In phase and quadrature multiplex
Multiple access methods	Frequency division multiple access	Time division multiple access	Code division multiple access

transmitters. However, with GPS, the satellites are all at roughly the same range, and the received signal levels normally do not vary greatly. (Exceptions occur when the signal from a given satellite is blocked momentarily by an aircraft wing tip, or, if on the ground, by tree foliage.) The choice of a specific family of codes for GPS that provide the desired code division multiple access performance is discussed in detail in later paragraphs.

An alternative multiple access technique considered was frequency division multiple access (FDMA). Frequency division multiple access of the satellite signals, which was subsequently selected for the GLONASS navigation satellites, has the advantage that the civil signals can be truly uncorrelated by offsetting the carriers in frequency by the bandwidth of the civil PN code. However, this approach occupies a larger bandwidth for a given code bandwidth, a disadvantage that the GLONASS developers diminished by operating the civil signal at roughly half the clock rate of the GPS signal. The GLONASS civil signal operates with a single 511-bit length PN code at 511 kbps, and spaces the carriers in frequency by 562.5 kHz.*[5] GLONASS is discussed more extensively in a later chapter. Decreasing the C/A-code clock rate for the same power flux density on the ground would provide a somewhat lesser accuracy. The other aspect of the frequency division approach felt to be a disadvantage was that the user receiver would have to operate with several frequency offsets if many satellites were to be tracked simultaneously. It was believed that the frequency division multiple access operation had a cost implementation disadvantage for the state of the art at that time (1973–1974).

B. GPS Radio Frequency Selection and Signal Characteristics

During the design phase of the GPS system, various frequency bands were considered for the GPS signal. Although a strict optimization of the frequency is not meaningful because only certain frequency bands could be made available, several considerations were important in selecting the GPS frequency band. Some of these are noted in Table 3.

The use of L-band gives acceptable received signal power with reasonable satellite transmit power levels and Earth coverage satellite antenna patterns, whereas the C-band path loss is roughly 10 dB higher because the path loss is proportional to f^2 for an omnidirectional receive antenna and fixed transmit antenna beamwidth and range. The large ionospheric delay and fluctuation in delay weighs against uhf as does the difficulty in obtaining two large ($\approx$20 MHz) bandwidth frequency assignments in the uhf band (two frequency bands are necessary for ionospheric correction). Thus, L-band was selected, and dual frequencies permit ionospheric group delay measurements. The signal bandwidths at both center frequencies are 20 MHz.

1. *Global Positioning System Signal Characteristics*

Thus, the GPS signal consists of two components, Link 1 or L_1, at a center frequency of 1575.42 MHz and Link 2, L_2, at a center frequency of 1227.6 MHz.

*The center frequencies of the channels are $1602 + 0.5625\,n$ MHz, where $n = 0, 1, 2, \ldots, 24$. See Chapter 9 in the companion volume.

Table 3 Global positioning system transmission frequency band selection considerations

Performance parameter	uhf ≈400 MHz	*L*-band (1–2 GHz)	*C*-band (4–6 GHz)
Path loss for omnidirectional receive antenna-loss $\sim f^2$	Path loss lowest of the three	Acceptable	Path loss ≈ 10 dB larger than at *L*-band
Ionospheric group delay, $\Delta R \sim 1/f^2$	Large group delay, 20–1500 ns	Group delay 2–150 ns at 1.5 GHz	Group delay ≈ 0–15 ns
Other considerations	Galactic noise ≈ 150°K at 400 MHz	—	Rainfall/atmospheric attenuation can be significant in 4–6 GHz band 0.1 to 1 dB/km at 100 mm of rain/hour

Each of the center frequencies is a coherently selected multiple of a 10.23 MHz master clock. In particular the link frequencies are the following:

$$
\begin{aligned}
L_1 &= 1575.42 \text{ MHz} = 154 \times 10.23 \text{ MHz} \\
L_2 &= 1227.6 \text{ MHz} = 120 \times 10.23 \text{ MHz}
\end{aligned}
\tag{7}
$$

Similarly, all of the signal clock rates for the codes, radio frequency carriers, and a 50 bps navigation data stream are coherently related.

The frequency separation between L_1 and L_2 is 347.82 MHz or 28.3%, and it is sufficient to permit accurate dual-frequency estimation of the ionospheric group delay. (The ratio of $L_1/L_2 = 77/60 = 1.2833$.) The ionospheric group delay varies approximately as the inverse square of frequency, and thus measurement at two frequencies permits calculation of the ionospheric delay. The ionospheric group delay correction is obtained by subtracting the total L_1 group delay τ_{GDL_l} from the total L_2 group delay τ_{GDL_2} in order to cancel the true pseudorange delay. This difference $\Delta\tau$ is then (neglecting random noise for the moment) the following:

$$
\Delta\tau = \tau_{GDL_2} - \tau_{GDL_1} = \frac{A}{f_{L_2}^2} - \frac{A}{f_{L_1}^2} = \frac{A}{f_{L_1}^2}\frac{1}{1.54573} = \frac{\tau_{\text{iono}}}{1.54573}
\tag{8}
$$

or

$$
\tau_{\text{iono}} = A/f_{L_1}^2 = 1.54573\ \Delta\tau
$$

where τ_{iono} is the ionospheric group delay at L_1, and $\Delta\tau$ is the measurable difference between total propagation delays at L_1 and L_2. Thus, the frequencies L_1 and L_2 are separated far enough in frequency so that the ratio is only a factor 1.54573. Ionospheric effects and models are discussed both in Chapter 4 and in considerable detail in Chapter 12, this volume.

As discussed in Chapter 18, the relativistic effects are not negligible in GPS but are partially compensated for in the satellite by offsetting the 10.23 MHz master clock rate to a slightly lower number before launch. As the signal approaches the Earth from the satellite, the frequency increases slightly because of relativity by approximately the same factor as the offset, and for a stationary user on the Earth's surface, the GPS signal clock appears to have a frequency very close to the desired 10.23 MHz. Henceforth, when reference is made to the desired 10.23 MHz, the frequency will always be this slightly offset frequency as far as the satellite clocks are concerned when observed prior to launch. The actual frequency of the satellite clocks before launch is 10,229,999.995453 MHz or an offset of $\Delta f = 4.57$ mHz below 10.23 MHz. The fractional frequency offset is $-\ 4.46 \times 10^{-10}$(see Ref. 6).

The L_1 signal is modulated by both a 10.23 MHz clock rate precision P signal and by a 1.023 MHz civil C/A signal to be used by the civil user. The transmitted signal spectra for both L_1 and L_2 are shown in Fig. 7. The binary modulating signals are formed by a P code or a C/A code that is modulo-2 added to the 50 bps binary data D, to form* P⊕D and C/A ⊕ D, respectively. The P code also can be converted to a secure antispoof Y code at the same clock rate, and is labeled the P(Y) code. The L_1 signal has an in-phase component of its carrier that is modulated by the P signal, P⊕D, and a quadrature (within ±100 m rad) carrier component that is modulated by C/A⊕D. The peak power spectral density of the C/A signal exceeds that of the P code at L_1 by approximately 13 dB because it is nominally 3 dB stronger and has 1/10 the chip rate and bandwidth. The in-phase and quadrature waveforms and phasor diagram of the L_1 signal are shown in Fig. 8.

The L_2 signal is biphase-modulated by either the P code or the C/A code. Normal operation would provide P- or Y-code [labeled P(Y)] modulation on the L_2 signal. There may or may not be data modulation on L_2 dependent on ground command.

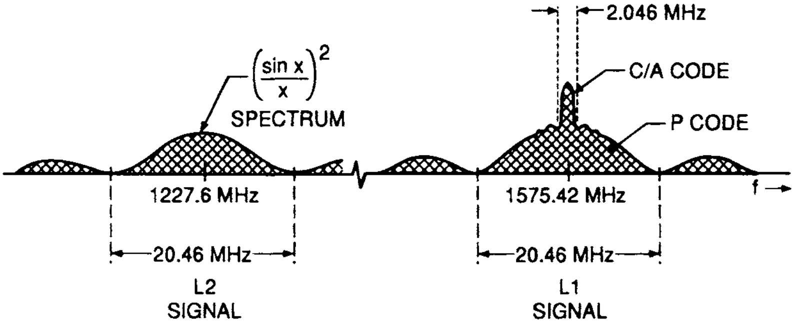


Fig. 7 GPS power spectral density.

*The symbol ⊕ stands for modulo-2 addition of 0, 1 numbers, which is equivalent to multiplication of +1, −1 numbers, respectively.

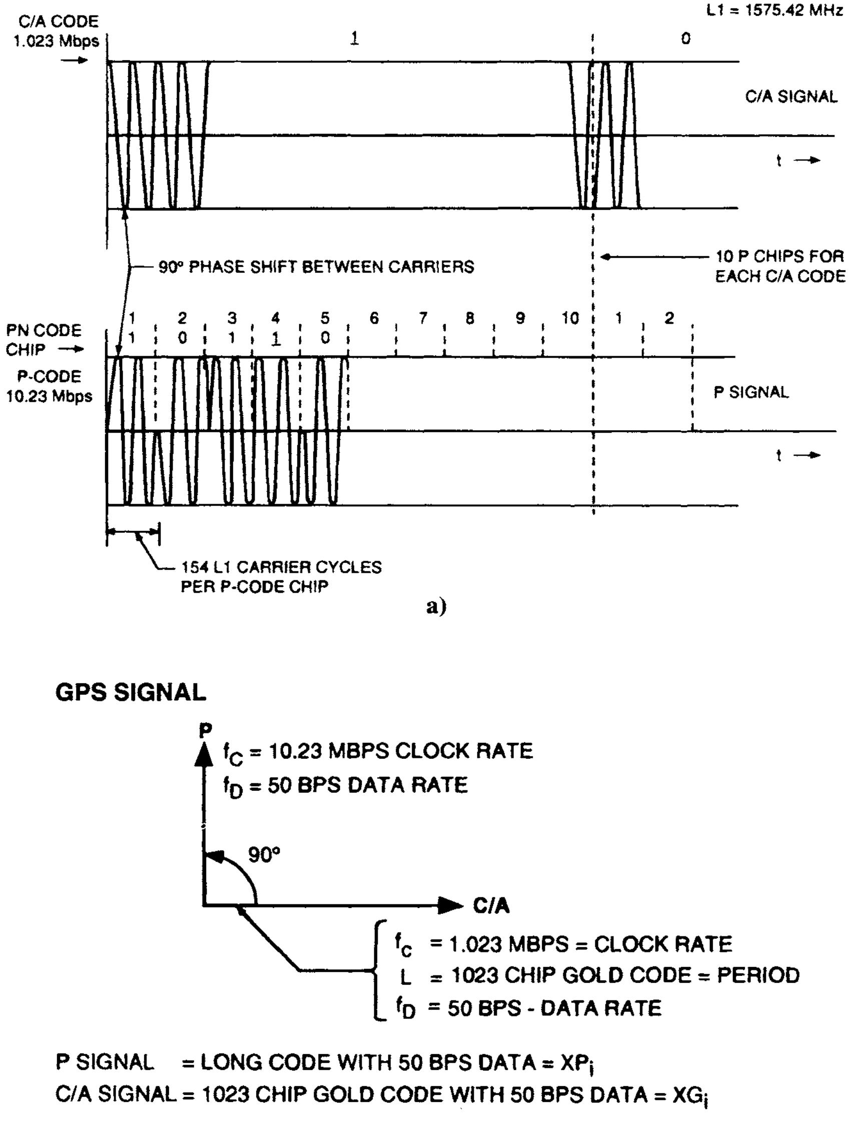

L1 = 1575.42 MHz
C/A CODE
1.023 Mbps
1
0
C/A SIGNAL
t →
90° PHASE SHIFT BETWEEN CARRIERS
10 P CHIPS FOR
EACH C/A CODE
PN CODE
CHIP →
P-CODE
10.23 Mbps
1 2 3 4 5 6 7 8 9 10 1 2
1 0 1 1 0
P SIGNAL
t →
154 L1 CARRIER CYCLES
PER P-CODE CHIP
a)
GPS SIGNAL
P
f_C = 10.23 MBPS CLOCK RATE
f_D = 50 BPS DATA RATE
90°
C/A
f_C = 1.023 MBPS = CLOCK RATE
L = 1023 CHIP GOLD CODE = PERIOD
f_D = 50 BPS - DATA RATE
P SIGNAL = LONG CODE WITH 50 BPS DATA = XP_i
C/A SIGNAL = 1023 CHIP GOLD CODE WITH 50 BPS DATA = XG_i
b)

Fig. 8 GPS L_1 signal waveform and phasor diagram. The P code for satellite i is labeled XP_i, and the C/A code for satellite i is labeled XG_i. a) Radio frequency waveforms for the P signal and C/A signal (carrier not to scale). b) Phasor diagram.

The GPS satellite can also transmit a third L-band carrier modulated by the C/A code at an L_3 frequency of 1381.05 MHz = 135 × 10.23 MHz. This signal is utilized only in a time-gated mode for a Nudet (Nuclear Detonation) Detection System (NDS) and is not used in the GPS navigation receiver.

C. Detailed Signal Structure

The L_1 signal contains both in-phase and quadrature signals. The signal transmitted (see Fig. 8) by the satellite i is then as follows:

$$S_{L_{1i}}(t) = \sqrt{2P_c}XG_i(t)D_i(t)\cos(\omega_1 t + \phi) + \sqrt{2P_p}XP_i(t)D_i(t)\sin(\omega_1 t + \phi) \tag{9}$$

where ω_1 is the L_1 frequency as defined above, ϕ represents a small phase noise and oscillator drift component and P_c and P_p are the C/A and P signal powers, respectively. Oscillator frequency stability is obtained using redundant cesium and rubidium atomic frequency standards. (The first satellite in the GPS series NTS-2, had a long-term clock stability better than 2×10^{-13} and later satellites have improved stability $\approx 3 \times 10^{-14}$). The P code $XP_i(t)$, is a ±1 pseudorandom sequence with a clock rate of 10.23 Mbps, and a period of exactly 1 week. Each satellite i transmits unique C/A and P codes. The binary data $D_i(t)$, also has amplitude ±1 at 50 bps and has a 6 s subframe and a 30 s frame period. The C/A code XG_i is a unique Gold code of period 1023 bits and has a clock rate of 1.023 Mbps. Thus, the C/A code has a period of 1 ms.

In GPS, the C/A-code strength is nominally 3 dB stronger than the P code on L_1. As already mentioned above, the code clocks and transmitted radio frequencies are all coherently derived from the same on-board satellite frequency standard. The rms clock transition time difference between the C/A and P code clocks is required to be less than 5 ns. Both C/A and P codes are of a class called product codes; i.e., each is the product of two different code generators clocked at the same rate where the delay between the two code generators defines the satellite code i (see Fig. 9). The specific component codes forming the product code for P and C/A are quite different, but the principle is similar. The clock interval for the C/A code $T_{cc} = 10T_c$ where T_c is the P-code clock interval in the figure.

1. P Code—Precision Code

The P code for satellite i is the product of 2 PN codes, $X1(t)$ and $X2(t + n_iT)$, where $X1$ has a period of 1.5 s or 15,345,000 chips, and $X2$ has a period of 15,345,037 or 37 chips longer. Both sequences are reset to begin the week at the same epoch time. Both $X1$ and $X2$ are clocked in phase at a chip rate $f_c = 1/T_c = 10.23$ MHz. Thus, the P-code is a product code of the following form:

$$XP_i(t) = X1(t)X2(t + n_iT), \quad 0 \le n_i \le 36 \tag{10}$$

where $X1(t)$ and $X2(t)$ are binary codes of value ±1 and $XP_i(t)$ is reset at the beginning of the week. The delay between $X1(t)$ and $X2(t)$ is n_i code clock intervals of T_c s each (see Fig. 10). The $X1$ and $X2$ codes are each generated as the products of two different pairs of 12-stage linear feedback shift registers)

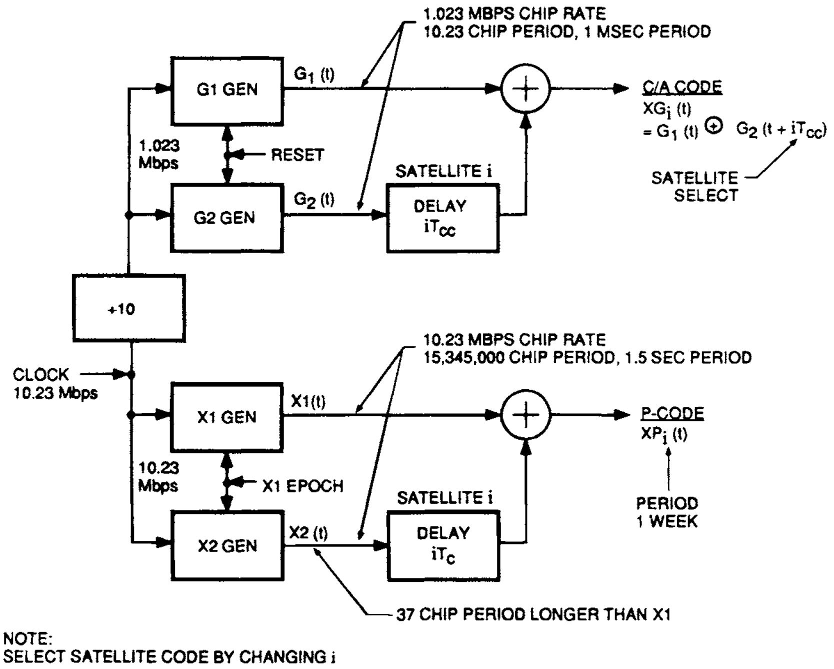


Fig. 9 GPS code generators for satellite *i*. Both the C/A- and P-codes are product codes.

$X1A$ and $X1B$ and $X2A$ and $X2B$ with polynomials specified in the GPS-ICD-200[6] as follows:

$$
\begin{aligned}
&X1A\text{: } 1 + X^6 + X^8 + X^{11} + X^{12} \\
&X1B\text{: } 1 + X + X^2 + X^5 + X^8 + X^9 + X^{10} + X^{11} + X^{12} \\
&X2A\text{: } 1 + X + X^3 + X^4 + X^5 + X^7 + X^8 + X^9 + X^{10} + X^{11} + X^{12} \\
&X2B\text{: } 1 + X + X^3 + X^4 + X^8 + X^9 + X^{12}
\end{aligned}
\tag{11}
$$

See the Appendix for a discussion of code polynomials, Galois fields, and shift registers.

For now, suffice it to say that these polynomials give the feedback tap positions of the 12-stage shift registers, $X1A$, $X2A$, $X1B$, $X2B$, of Fig. 10. Recall the introductory discussion of Fig. 5. The $X1A$ and $X1B$ codes have different relatively prime periods as do the $X2A$ and $X2B$ codes. A 12-stage maximal length shift register produces a code period of $2^{12} - 1 = 4095$. If two code generators are short cycled to give relatively prime periods less than or equal to 4095, then the product code can have a period in the vicinity of 1.6×10^7, the product of the two periods. For GPS, the two product codes have been short cycled to relatively prime periods of 15,345,000 and 15,345,037 for the $X1$ and $X2$ respectively. Likewise, the product of $X1$ and $X2$ codes generates a new code that has a period that is the product of the periods, unless it is short cycled.

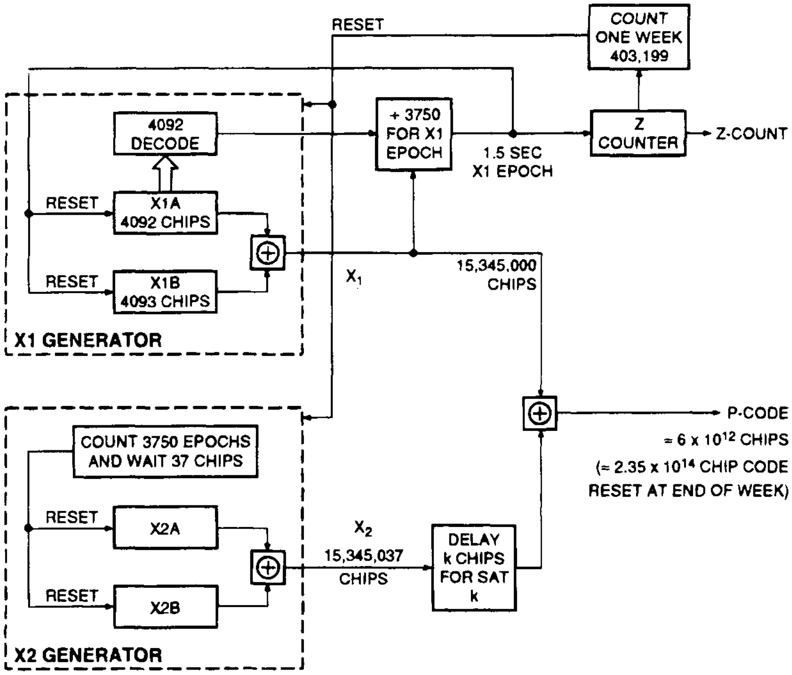


Fig. 10 Simplified P-code generator block diagram.

The product of $X1$ and $X2$ codes clocked together act somewhat like two gears with a number of teeth on each gear corresponding to the periods of $X1$ and $X2$ as shown in Fig. 11. If we imagine that the teeth of both gears are coded in black and white according to the respective binary chips in the PN sequence, then for relatively prime code lengths for $X1$ and $X2$ the period of this product code is equal to the product of the two individual codes periods (the number of teeth on each gear wheel). Each satellite has a unique code offset n_iT between code $X1$ and code $X2$, which makes the P-code unique as well. The increase in code period for $X2$ by 37 relative to $X1$ allows the values of n_i to range over 0–36 without having any significant segment of a P code of one satellite match that of another. Thus, we have 37 different pseudorandom P codes.

For a different view of the P code, note that the period of a product of $X1$ and $X2$ codes, each of which has a relatively prime period, is the product of the periods; i.e. $(15{,}345{,}000) \times (15{,}345{,}037) = 2.35469592765 \times 10^{14}$. Thus, if the P-code were allowed to continue without being reset, each P code would continue without repetition for slightly more than 38 weeks. In effect, this overall period has been subdivided so that each of 37 possible GPS satellites or ground transmitters (pseudolites) gets a 1-week period code, which is nonoverlapping with that of any other satellite.

A long period code such as the P code is difficult to acquire without acquisition aids. For example, a receiver correlator must be timed to within roughly one P-code chip or roughly 0.1 μs and clocked in synchronism in order to correlate at

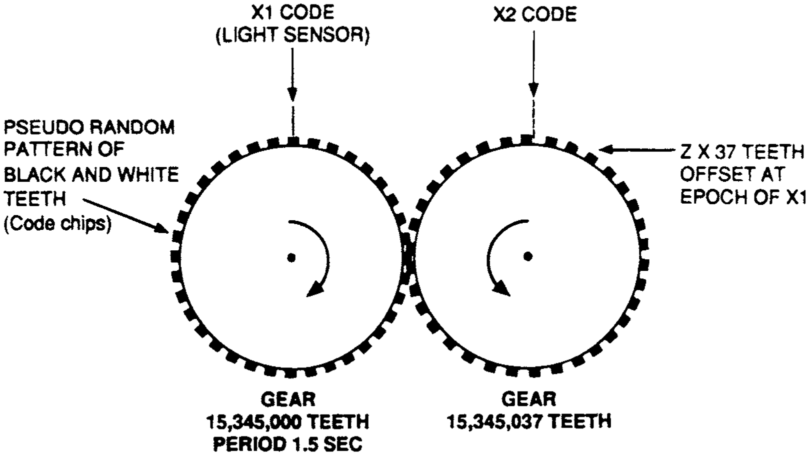


Fig. 11 P-code–subcode characteristics as represented pictorially by two gear wheels with light sensors for both gear wheels.

all. Note that the period of the $X1$ code is exactly 1.5 s, i.e., $1.5 \times 10.23 \times 10^6 = 15{,}345{,}000$, and there are this same number of code chip time bins to search.

A timing Z-count is defined in Fig. 12 as the number of 1.5 s $X1$ epochs since the beginning of the week. There are four $X1$ epochs per data subframe of 6 s. To help the receiver to acquire the long period P code, the 50 bps datastream contains an updated handover-word (HOW) for each 6-s subframe. The HOW when multiplied by 4, equals the Z-count at the beginning of the next 6-s subframe. Thus, if we have acquired timing from the relatively short C/A code and know the subframe epoch times and the HOW words, we can instantly acquire the P-code at the next subframe epoch. Figure 13 summarizes the timing relationships between $X1$, $X2$ epochs, and the Z-count and HOW words.

2. *Antispoof P(Y) Code, Nonstandard Codes, and Selective Availability*

The P code is a long, 1-week period code; however, it is published in GPS-ICD[6] and is available to potential spoofers or jammers. (A spoofer generates a signal that mimics the GPS signal and attempts to cause the receiver to track the wrong signal.) For this reason, the GPS system has the option to replace the P code with a secure Y code available only to authorized U.S. Government users. The Y code is employed when the "antispoof" or AS mode of operation is activated. The Y code is a secure version of the published P code that operates at the same clock rate as the P code, but has a code available only to authorized users. The main purpose of the Y code is to assure that an opponent cannot spoof the Y-code signal by generating a Y-code replica.

Nonstandard C/A and Y codes (NSC and NSY codes) are used in place of the C/A and P(Y) codes to protect the user from a malfunction in the spacecraft. They are only used for a malfunctioning satellite. These codes are not used in navigation receivers.

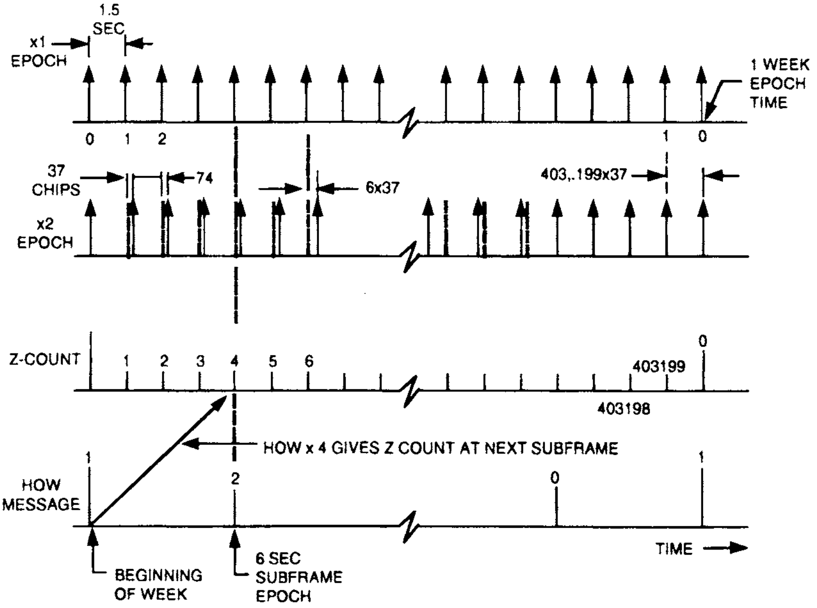


Fig. 12 Timing diagram for the P-code components X_1, X_2, and the Z-count and HOW message relationship (not to scale). The HOW message is carried in the 50-bps datastream.

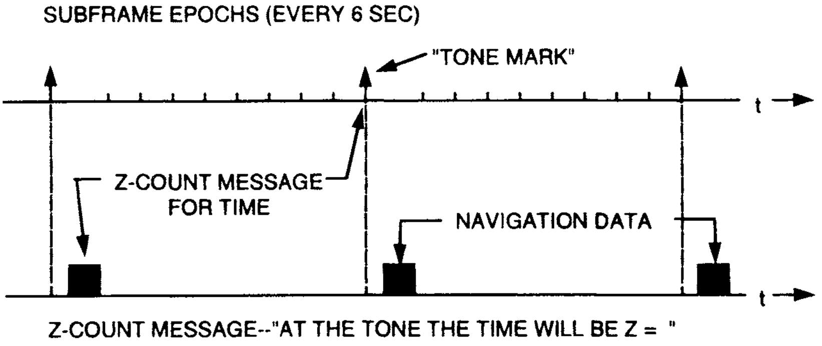

SUBFRAME EPOCHS (EVERY 6 SEC)
"TONE MARK"
t
Z-COUNT MESSAGE
FOR TIME
NAVIGATION DATA
t
Z-COUNT MESSAGE--"AT THE TONE THE TIME WILL BE Z = "

Fig. 13 GPS received signal time and the Z-count navigation data that are used to help acquire the P-code once the C/A code is acquired. The Z-count also aids in time ambiguity resolution for the C/A code.

Selective availability is a purposeful degradation of the GPS signal by the U.S. Government that can be imposed to restrict the full accuracy of the GPS system to authorized military users. Selective Availability (SA) is discussed in a later chapter in this volume. RMS position accuracy with SA is less than or equal to 100 m. SA purposely dithers the GPS clock in a pseudorandom manner. The clock dither has been reported by Allan and Dewey[7] to have a decorrelation time of 300–400 s. For observation times shorter than 300 s the clock dither can be modeled as a random walk phase modulation; for longer observation times it can be modeled as white noise phase modulation. As discussed in Chapter 2, this volume, recommendations have been made to turn selective availability to zero.

3. *Coarse/Acquisition (C/A) Code—Civil GPS Signal*

The C/A code for the civil user is a relatively short code with a period $2^{10} - 1 = 1023$ bits or 1-ms duration at a 1.023 Mbps bit rate. The code period is purposely selected to be relatively short so as to permit rapid acquisition. That is, there are only 1023 code chip time bins to search. The C/A codes are selected to provide good multiple access properties for their period. The C/A codes for the various satellites are taken from a family of codes known as Gold codes that are formed by the product of two equal period 1023 bit PN codes $G1(t)$ and $G2(t)$ (see Appendix and Ref. 8). Thus, this product code is also of 1023 bit period and is represented as follows:

$$XG(t) = G1(t)G2[t + N_i(10T_c)] \tag{12}$$

where N_i determines the phase offset in chips between $G1$ and $G2$. Note that C/A-code chip has duration $10T_c$ s where T_c is the P-code chip interval. There are 1023 different offsets N_i, and hence 1023 different codes of this form.* Each code $G1$, $G2$ is generated by a maximal-length linear shift register of 10 stages. The $G1$ and $G2$ shift registers, are set to the "all ones" state in synchronism with the $X1$ epoch. The tap positions are specified by the generator polynomial for the two codes:

$$\begin{aligned} G1&: G_1(X) = 1 + X^3 + X^{10} \\ G2&: G_2(X) = 1 + X^2 + X^3 + X^6 + X^8 + X^9 + X^{10} \end{aligned} \tag{13}$$

Because each Gold code has a 1-ms period, there are 20 C/A-code epochs for every databit. The 50-bps data clock is synchronous with both the C/A epochs and the 1.5 *s* $X1$ epochs of the P code. Figure 14 shows a simplified block diagram of the C/A-code generator. The unit is comprised of two 10-stage feedback shift registers clocked at 1.023 Mbps having feedback taps at stages 3 and 10 for $G1$ and at 2, 3, 6, 8, 9, 10 for $G2$, as indicated by the polynomials of (13). The various delay offsets are generated by tapping off at appropriate points on the $G2$ register and modulo-2 adding the two sequences together to get the desired delayed version of the $G2$ sequence using the so-called "cycle-and-add" property of the linear maximal length shift register (LMLSR). Maximal length shift register codes have the property that the addition of two time offset ("cycled") versions of the same code gives a shifted

*There actually are 1025 different Gold codes of this period and family. The codes, $G1(t)$ and $G2(t)$, by themselves, are the other two codes.

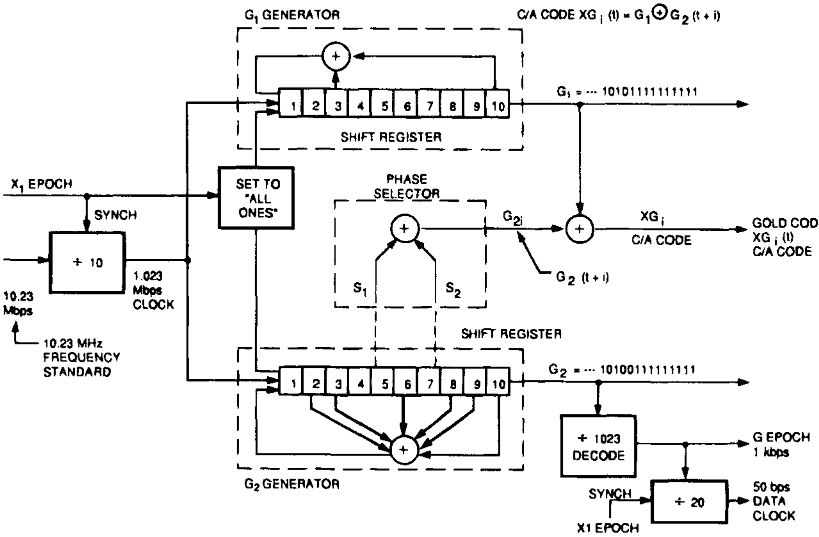


Fig. 14 C/A-code generator block diagram showing the 1-ms *G* epoch and data clock generators that are all in synchronism.

version of the same code; hence the name "cycle and add" property. This property is discussed later in the chapter. The code tap positions for various codes are given in Table 4. Note that there are 45 possible tap positions in Fig. 14, but as shown in Table 4, only 37 codes are defined in GPS-ICD-200.[6] The general relationship between code taps and code phase is analyzed in the Appendix. Epochs of G code occur at 1 Kbps and are divided down by 20 to get the 50 bps data clock. All clocks are in phase synchronism with the $X1$ clock, as shown in Fig. 14.

The recursive equations for the $G1$ and $G2$ sequences that correspond to the $G1$, $G2$ polynomials of (2–8) are as follows:

$$
\begin{aligned}
G1(i) &= G1(i-10) \oplus G1(i-3) \\
G2(i) &= G2(i-10) \oplus G2(i-9) \oplus G2(i-8) \\
&\quad \oplus G2(i-6) \oplus G2(i-3) \oplus G2(i-2)
\end{aligned}
\tag{14}
$$

Table 4 gives the first bits of each of the 37 C/A codes in octal form. For reference, the first 31 bits of the first Gold code for SV#1 are {1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1}.

Notice that the first 10 bits of this code match exactly those shown in Table 4.

It is also possible to generate each Gold code with a single 20-stage shift register (not maximal length) by simply using a code generator that corresponds to a polynomial that is the product of the $G1(x)$ and $G2(x)$ polynomials (modulo-2). This code generator still produces codes of length 1023. Different codes are formed by starting the shift register in the correct state. However, that form does not allow one generator to generate all codes easily. It is also possible simply to delay one code generator, the $G2$ generator, relative to $G1$ simply by changing the starting state of $G2$.

Table 4 GPS code phase assignments for various spacecraft ID numbers (taken from GPS-ICD-200)

SV ID No.	GPS PRN signal No.	Code phase selection, C/A, ($G2_+$)	$X2_+$	Code delay chips C/A	P	First 10 chips octal[a] C/A	First twelve chips octal P
1	1	$2\oplus6$	1	5	1	1440	4444
2	2	$3\oplus7$	2	6	2	1620	4000
3	3	$4\oplus8$	3	7	3	1710	4222
4	4	$5\oplus9$	4	8	4	1744	4333
5	5	$1\oplus9$	5	17	5	1133	4377
6	6	$2\oplus10$	6	18	6	1455	4355
7	7	$1\oplus8$	7	139	7	1131	4344
8	8	$2\oplus9$	8	140	8	1454	4340
9	9	$3\oplus10$	9	141	9	1626	4342
10	10	$2\oplus3$	10	251	10	1504	4343
11	11	$3\oplus4$	11	252	11	1642	4343
12	12	$5\oplus6$	12	254	12	1750	4343
13	13	$6\oplus7$	13	255	13	1764	4343
14	14	$7\oplus8$	14	256	14	1772	4343
15	15	$8\oplus9$	15	257	15	1775	4343
16	16	$9\oplus10$	16	258	16	1776	4343
17	17	$1\oplus4$	17	469	17	1156	4343
18	18	$2\oplus5$	18	470	18	1467	4343
19	19	$3\oplus6$	19	471	19	1633	4343
20	20	$4\oplus7$	20	472	20	1715	4343
21	21	$5\oplus8$	21	473	21	1746	4343
22	22	$6\oplus9$	22	474	22	1763	4343
23	23	$1\oplus3$	23	509	23	1063	4343
24	24	$4\oplus6$	24	512	24	1706	4343
25	25	$5\oplus7$	25	513	25	1743	4343
26	26	$6\oplus8$	26	514	26	1761	4343
27	27	$7\oplus9$	27	515	27	1770	4343
28	28	$8\oplus10$	28	516	28	1774	4343
29	29	$1\oplus6$	29	859	29	1127	4343
30	30	$2\oplus7$	30	860	30	1453	4343
31	31	$3\oplus8$	31	861	31	1625	4343
32	32	$4\oplus9$	32	862	32	1712	4343
——[c]	33	$5\oplus10$	33	863	33	1745	4343
——[c]	34[b]	$4\oplus10$	34	950	34	1713	4343
——[c]	35	$1\oplus7$	35	947	35	1134	4343
——[c]	36	$2\oplus8$	36	948	36	1456	4343
——[c]	37[b]	$4\oplus10$	37	950	37	1713	4343

[a] In the octal notation for the first 10 chips of the C/A code as shown in this column the first digit (1) represents a "1" for the first chip and the last three digits are the conventional octal representation of the remaining 9 chips. (For example, the first 10 chips of the C/A code for PRN Signal Assembly No. 1 are: 1100100000).

[b] C/A codes 34 and 37 are common

[c] PRN sequences 33 through 37 are reserved for other uses (e.g. ground transmitters).

$\oplus$ = "exclusive or"

We might ask why not simply take 37 different maximal length shift register codes and use them in place of the Gold codes? After all, 2 of the Gold codes of this family of length 1023 are the maximal length codes themselves. The answer is that the other maximal length codes of length 1023 would not guarantee uniformly low cross-correlation sidelobes for all other needed satellite codes and all possible delay offsets.

4. *L_2 Signal*

The L_2 signal is biphase-modulated by either the P code or the C/A code, as selected by ground command. The same 50 bps datastream modulates the L_2 carrier as transmitted on L_1. Thus, the L_2 signal is represented in the normal P format as follows:

$$S_{L_{2_i}}(t) = \sqrt{2P_2}XP_i(t)D_i(t)\cos(\omega_2 t + \phi_2) \tag{15}$$

where $\sqrt{2P_2}$ represents the L_2 signal amplitude at the satellite, $XP_i(t)$ is the P code for the ith satellite, which is clocked in synchronism with the L_1 codes. As with the L_1 signal, both L_2 carrier and code are synchronous with one another. The L_2 signal can also be modulated with the P-code without the data. This feature permits the precision receiver tracking loops to be reduced further in IF bandwidth, and thereby can improve the noise/interference performance.

Because the L_2 signal is biphase-modulated, it is possible to recover the L_2 carrier without knowledge of the P code by simply squaring the signal or cross-correlating L_1 with L_2 with a delay offset that matches the L_1 to L_2 ionospheric delay difference. These types of codeless recovery of the L_2 carrier can then be used in estimating the ionosphere delay (see Chapter 4, this volume). There is added noise degradation in this codeless carrier recovery because of the nonlinear squaring operation or noisy cross-correlation operation. However, the information bandwidth of the ionosphere is sufficiently small that if we have already tracked the L_1 code, the noise bandwidth required to track the L_2–L_1 ionospheric difference is also very small, and noise effects can be kept small by narrow bandwidth filtering.

5. *GPS Data Format*

Table 5 summarizes the signal and data characteristics just discussed. There are five subframes of 6 s each for a total frame period of 30 s. One of the key points to be made in the signal structure discussion is that acquisition by a receiver of the relatively short period C/A code plus the recovery of a single full subframe of data permits us to acquire the P code with minimal or zero search. Knowledge of the C/A epoch plus the data subframe epoch and the HOW word gives the exact phasing of the P(Y) code. Navigation solutions require, as a minimum, reception of data subframes 1, 2, 3 containing clock-correction and ephemeris data and, in general, require reception of a full 30- s frame of data. The navigation data are discussed in detail in the next chapter.

6. *Codes for GPS Augmentation*

The GPS satellites are augmented at times by pseudolites or ground transmitters that may transmit frequency offset or a low-duty factor pulsed GPS-type signal

Table 5 Summary of GPS signal parameters and data formats

Parameter	C/A Signal	P Signal
Code clock (chip) rate	1.023 Mbps	10.23 Mbps
Code period	1023	= 6 x 10^{12}; 1 week
Data rate	50 bps	50 bps
Transmission frequency	L_1	L_1, L_2

Data formal—frame and subframe structure

Subframe No.	Ten, 30-bit words, 6-s subframe			
1	TLM	HOW	Block 1—Clock correction + satellite quality	1-Frame 30 s 1500 bits
2	TLM	HOW	Block 2—Ephemeris	
3	TLM	HOW	Block 3—Ephemeris continued	
4	TLM	HOW	Block 4—Almanac + ionosphere + UTC correction	
5	TLM	HOW	Block 5—Almanac—(25 frames for complete almanac)	

Each Telemetry (TLM) word contains an 8-bit Barker word for synchronization. The Handover Word (HOW) contains a 17-bit Z-count for handover from the C/A code to the P code.

and use codes different from those employed in the GPS satellites to avoid confusion. GPS pseudolites are discussed in detail in a later chapter.

Also planned is the augmentation of the GPS satellites with geostationary overlay satellites using different Gold codes. These signals might be generated either by the satellite or generated and synchronized on the ground and broadcast to the GPS users via geostationary satellite transponders. Table 6 lists the Gold codes selected by INMARSAT for future transmission by the INMARSAT satellites.[9] Table 6a lists the codes for the Wide Area Augmentation System (WAAS) of the Federal Aviation Administration. GPS augmentation is also discussed in later chapters on the GPS wide area differential GPS (WADGPS) and Wide Area Augmentation System in Volume II.

III. GPS Radio Frequency Receive GPS Power Levels and Signal-to-Noise Ratios

A. GPS Radio Frequency Signal Levels and Power Spectra

The minimum specified received signal strength for a user receiver employing a 0 dBIC antenna is given below in Table 7 for a satellite at elevation angles above 5 deg. As shown in the next subsection, the actual minimum varies with elevation angle to the satellite because of the shaped satellite antenna pattern. The signal power spectral densities for the P and C/A signal components are shown in Fig. 15. Figure 15 also shows the measured radiofrequency power spectral density of the L_1 signal. Note the narrowband high-power density C/A signal in the center of the signal spectrum. Note that these spectra are

Table 6 Final INMARSAT C/A-code selection

Order	PRN	Delay	Initial *G*2 state (octal)
1	201	145	1106
2	205	235	1617
3	208	657	717
4	206	235	1076
5	202	175	1241
6	207	886	1764
7	209	634	1532
8	211	355	341

Table 6a Wide Area Augmentation System (WAAS) PRN ranging C/A codes (Note that these codes include the INMARSAT codes of Table 6)

PRN Code #	Display (Chips)	First 10 WAAS Chips (Octal)[a]
115	145	0671
116	175	0536
117	52	1510
118	21	1545
119	237	0160
120	235	0701
121	886	0013
122	657	1060
123	634	0245
124	762	0527
125	355	1436
126	1012	1226
127	176	1257
128	603	0046
129	130	1071
130	359	0561
131	595	1037
132	68	0770
133	386	1327

[a] The first digit represents a 0 or 1 in the first chip. The next three digits are the octal representation of the remaining nine chips.

Table 7 GPS minimum received signal power levels at output of a 0 dBIC antenna with right-hand circular polarization[a]

	GPS signal component (minimum strength) specified		Expected maximum does not exceed this level with 0.6 dB atmospheric loss	
Link	P	C/A	P	C/A
L_1	−163 dBW	−160 dBW	−155 dBW	−153 dBW
L_2	−166 dBW	−166 dBW	−158 dBW	−158 dBW

[a] The satellite is assumed to be at an elevation angle ≥ 5 deg.

the transmitted spectra. In normal operation, the thermal white noise of the receiver significantly exceeds the signal spectral density, and the signal is not visible using a spectrum analyzer. Recall that the maximum power spectral density for a pseudonoise signal with a continuous $[(\sin \pi f/f_c)/\pi f/f_c)]^2$ shaped spectrum is P_s/f_c. Thus, if $P_s = -160$ dBW, the maximum power density is −160 dBW − 60.1 dB = −220.1 dBW/Hz.

It is useful to compare this number with the recommended power flux density limit of the CCIR[1] cited earlier in this chapter in Sec. I.A. If an effective antenna aperture area* of $A = \lambda^2/4\pi = 2.8856 \times 10^{-3} \text{m}^2$ or -25.4 dBm^2 at 1.57542 GHz is assumed for a unity gain antenna, then the power flux density per Hz is $P_s/f_sA = -194.7$ dBW/Hz-m^2. The total flux density in a 4-kHz band is then $4 \times 10^3\ P_s/f_sA = -158.7$ dBW/m^2, which is within the level recommended by the CCIR[1] of −154 dBW/m^2.

B. Satellite Antenna Pattern

The radio frequency received signal levels are transmitted from the satellite by shaped pattern antennas (see Fig. 16) to compensate partially for the increased path loss to the user at low-elevation angles. The GPS Block II satellite antenna is an array of helices on the face of the satellite. The edge of Earth is approximately 13.87 deg off the satellite antenna boresight, i.e., the Earth subtends an angle of approximately 27.74 deg from the GPS satellite altitude. The satellite antenna pattern extends somewhat beyond the edge of the Earth, as shown in Fig. 16. Thus, even a GPS receiver in another satellite can receive signals from GPS satellites to perform satellite positioning, provided that it is not blocked by the Earth's shadow and is not too far off the main lobe of the satellite antenna pattern (see Fig. 17).

The transmitted signal from the satellite is right-hand circularly polarized with an ellipticity (offset from perfectly circular) no worse than 1.2 dB for L_1 and 3.2 dB for L_2 within an angle of ±14.3 deg from boresight.† Because the user antenna can be at various orientations with respect to the satellites, the satellite received power is specified under the following conditions; a) the signal is measured at the output of a 3 dBIC (isotropic) linearly polarized receiving antenna; b) the space

*The relationship between antenna gain G and effective aperture area A for an ideal lossless antenna is $G = 4\pi A/\lambda^2$.[10]

†NAVSTAR GPS Space Segment/Navigation User Segment Interface Control Document, ICD-GPS-200, IRN-200B-005, Dec. 16, 1991.

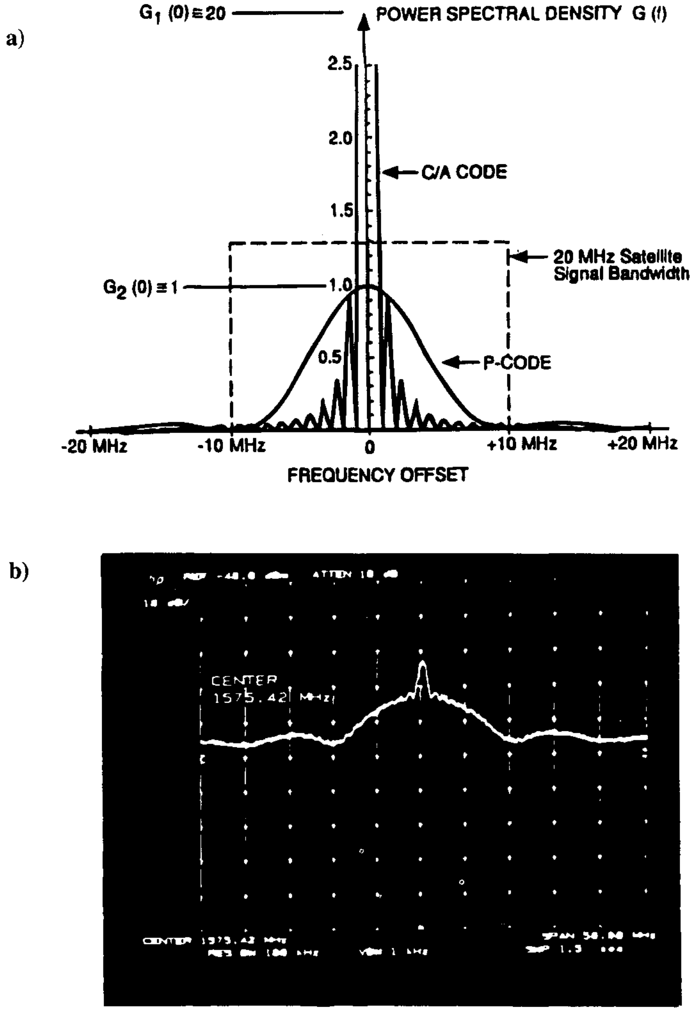


Fig. 15 Radio frequency spectrum plot and photograph of received L_1 carrier with C/A and P QPSK modulation; a) Power spectra of carriers with bit rates of 1.023 Mb/s and 10.23 mb/s. The ratio of C/A power to P-code signal power is 3 dB in this figure, b) Photograph of signal generated by Stanford Telecom GPS signal simulator 7200. Spectrum scales: horizontal, 5 MHz/division; vertical, 10 dB/division (courtesy Stanford Telecom).

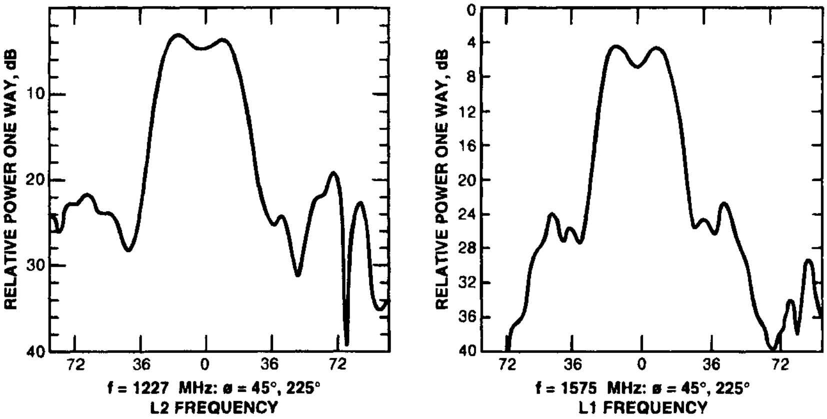


Fig. 16 Typical GPS satellite system transmit antenna patterns for the Block II satellites.

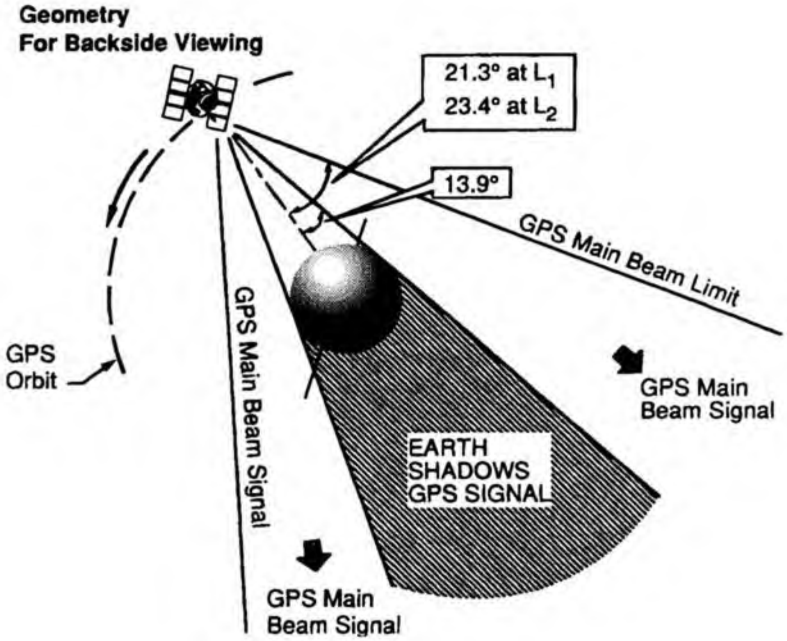


Fig. 17 GPS satellite main beam relative to Earth (not to scale). User satellites can navigate using GPS provided they are in the main beam of the GPS antenna but outside the Earth's shadow.

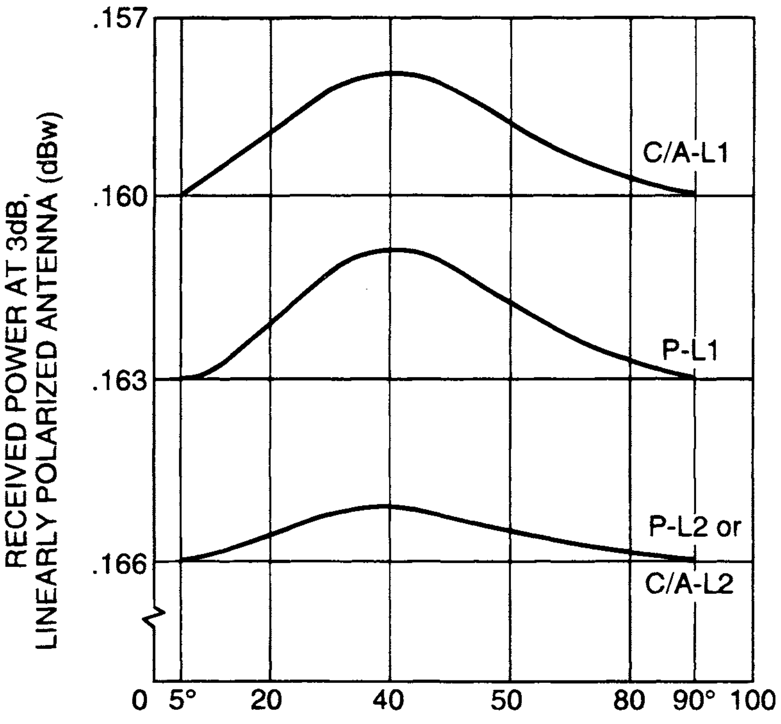


Fig. 18 GPS user received minimum signal levels vs elevation angle as stated in the GPS document GPS-ICD-200.[6]

vehicle is above a 5 deg elevation angle; c) the received signal levels are observed within the 20 MHz frequency allocation; d) the atmospheric path loss is 2.0 dB; and e) the space vehicle's attitude error is 0.5 deg (toward reducing signal level). The specified minimum received signal power vs. elevation angle for these conditions is shown in Fig. 18. Note that the specified received signal level peaks at 40 deg elevation angle at a level that is approximately 2 dB above the nominal − 160 dBW level for the C/A code on L_1. As shown in Chapter 13 this volume, atmospheric path loss is generally less than 2 dB, except at low-elevation angles. In addition, the satellites are designed so that these numbers are met at the end of the satellite's life. Beginning-of-life power levels are generally higher. Thus, these numbers are somewhat conservative.

C. Signal Specifications

1. Signal Correlation Loss

The GPS C/A and P(Y) signals are filtered to a bandwidth of 20.46 MHz. There are two such 20.46 MHz frequency bands centered at L_1 and L_2. The GPS space segment GPS-ICD-200[6] defines a maximum correlation loss of 1.0 dB

Chapter 4

GPS Navigation Data

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

GPS signal carries with it data from the satellite that the user receiver needs to solve for position, velocity, and time. This chapter describes these GPS navigation data in some detail and gives the background analysis useful in understanding their functions. The data formats correspond to the ICD-GPS-200.[1] The first section gives a complete overview of the entire navigation data format and overall frame structure. The second section describes in some detail the format and algorithms of ICD-GPS-200 for each of the subframes. The final sections present some of the physical and mathematical bases for the algorithms of Sec. II. Detailed discussion of the effects of relativity, the ionosphere, and troposphere is reserved for later chapters. A table of physical constants used in this chapter and elsewhere is given in the Appendix.

A. Overall Message Content of the Navigation Data

The 50 bits/s datastream provides the user with several key sets of data required to obtain a satisfactory navigation, geodetic survey, or time transfer solution. These navigation data are uploaded to each satellite by the GPS Control Segment (CS) for later broadcast to the user. Uploads occur once per day, or more often, if needed to keep the user range error (URE) within specification. The navigation data provide the information shown in Table 1. The form of the information is also described in the table.

1. Perturbing Factors in the Navigation Measurements

Various perturbations affect the relationship between measurements made on the received signal and the true range to the satellite from the user. Some of these are discussed in the preceding chapter. The remaining effects are discussed here and in the next chapter. Let us begin by defining (refer to Chapter 2, this volume) the "true" geometric pseudorange $\rho_{iT}(t)$, which assumes perfect

*Ph.D., Chairman of the Board.

Table 1 Requirements and characteristics of GPS navigation data

Requirement	Information provided by GPS navigation data
Precise satellite position at time of transmission	Satellite ephemeris using a modified Kepler model (sinusoidal perturbations) in an Earth–centered-inertial (ECI) coordinate frame with transformation to Earth-centered, Earth-fixed (ECEF) coordinates
Precise satellite time at time of transmission	Satellite clock error models and relativistic correction
P(Y) code[a] acquisition from C/A code[b]	A handover word (HOW) is transmitted that keeps track of the number of P(Y) code 1.5-s (*X*1 subsequence) periods thus far in the week. These data can aid in P(Y) code acquisition.
Select the best set of satellites for lowest appropriate GDOP[c] within elevation angle constraints (requires approximate knowledge of satellite position)	Moderate accuracy almanac that gives approximate position, time, and satellite health for the entire GPS constellation
Time transfer information	GPS time to universal coordinated time (UTC) time conversion data
Ionospheric corrections for single-frequency users	Approximate model of ionosphere vs time and user location
Quality of satellite signals/data	User range accuracy (URA)—a URA index *N* is transmitted that gives a quantized measure of space vehicle accuracy available to the civil (unauthorized) user

[a] P(Y) code, precise, antispoof code.
[b] C/A code, coarse/acquisition code.
[c] GDOP, Geometric Dilution of Precision.

knowledge of satellite clock time and position, assumes further the absence of any atmospheric propagation delay or relativistic effects, and uses a nonrotating ECI coordinate system. The "true" geometric pseudorange $\rho_{iT}(t)$ at time t for satellite i is then defined as follows:

$$\rho_{iT}(t) \triangleq = c(t_u - t_{si}) + cb_u = |\boldsymbol{x}_{si} - \boldsymbol{x}_u| + cb_u = |\boldsymbol{r}_{si} - \boldsymbol{r}_u| + cb_u = D_i + cb_u \tag{1}$$

where c is the speed of light, $\boldsymbol{x}_{si}(t - D_i(t)/c)$ is the true satellite position at the true time of transmission, $t - D_i(t)/c$, in ECI coordinates $\boldsymbol{x}_u(t)$ is the unknown user position also in ECI coordinates where t is the time of reception, and $t_u(t)/c$ is the user clock reading at the time of reception. Pseudorange is expressed in meters. User position is expressed either in Cartesian coordinates as $\boldsymbol{x}_u$ or in spherical coordinates as $\boldsymbol{r}_u$. The received satellite signal indicates the satellite clock time at the time of transmission* $t_{si}(t - D_i(t)/c)$, and $b_u(t)$ is the user clock time offset $b_u(t) = [t_u(t) - t_{\mathrm{GPS}}(t)]$ expressed in seconds. The geometric transit time of the signal is $D_i(t)$ expressed in meters. The coordinates can be rotated to ECEF coordinates, but the rotation of the Earth during the transit time $D_i(t)$ must also be taken into account.

The true pseudorange $\rho_{iT}(t)$ is not directly an observable, but instead must be observed with various perturbations. The measured pseudorange $\rho_i(t)$ is equal to the true pseudorange plus various perturbing factors, as shown below [refer again to Chapter 2, Eq. (2)].

$$\rho_i = \rho_{iT} + \Delta D_i - c\Delta b_i + c(\Delta T_i + \Delta I_i + \nu_i + \Delta\nu_i) \tag{1}$$

where

Δb_i = satellite bias clock error, s
ΔD_i = satellite position bias error effect on range, m
ν_i = receiver measurement noise error for satellite i, s
ΔI_i = ionospheric excess delay, s
ΔT_i = tropospheric excess delay, s
$\Delta \nu_i$ = relativistic time correction, s

To be precise one also must account for small second-order effects caused by motion of the satellite during the time interval caused by ionospheric and tropospheric excess delay when computing the satellite position at time of transmission.

Each of the perturbations in the pseudorange measurement equation above must be either estimated, measured, or computed using the navigation datastream or other information. Satellite position information itself, of course, must be estimated from the navigation datastream. Table 2 identifies which type of data is involved for each perturbing parameter. Figure 1 (modified from ICD-GPS-200)[1] illustrates in specific terms how each of the corrections is applied to the pseudorange measurements. Each of these parameters is discussed either in this chapter or in other chapters of this book. Note that the tropospheric corrections are not discussed in ICD-GPS-200.[1] However, tropospheric correction models are discussed in detail in Chapter 13, this volume.

B. Navigation Data Subframe, Frame, and Superframe

The navigation data are transmitted in a 50 bits/s stream that is modulo-2 added to the C/A and P(Y) codes on the L_1 frequency and may or may not be carried on the L_2 P(Y) code, depending on the satellite mode. The data bit stream is synchronous with the 1-kHz C/A code epochs. The databits are formatted into 30-bit words, and the words are grouped into subframes of 10 words that are 300 bits in length and 6 s in duration. Frames (or pages) consist of 5 subframes of 1500 bits and 30 s in duration, and a superframe consists of 25 frames and has a duration 12.5 min. The general frame and subframe format is shown in Fig. 2. Much of the data repeat every frame and some data; e.g. the 8-bit preamble repeat every subframe. Periodically, the navigation dataframes are updated. New navigation data set (4-h curve fit for normal operations) cutovers occur only on

*The time indicated is the "proper" time indicated by the satellite clock at the time of transmission.

Table 2 **Relationship between measured pseudorange and other parameters**

	Parameters	Source of information for parameter
True pseudorange	$\rho_{iT} = \lvert \boldsymbol{x}_{si} - \boldsymbol{x}_u \rvert + cb_u$	Estimated by the GPS receiver
Pseudorange measurement	$\rho_i(t)$	Measured by GPS navigation receiver
Satellite position for satellite *i*	$\boldsymbol{x}_{si}(t)$	Satellite position, calculated from the navigation data
User position	$\boldsymbol{x}_u(t)$	Unknown user position, to be estimated
Ionospheric excess path delay	$\Delta I_i(t)$	Measured using dual-frequency measurements, or for single-frequency user, modeled using navigation data
Tropospheric excess path delay	$\Delta T_i(t)$	Estimated using approximate equations of various levels of accuracy, some of which rely on pressure and humidity measurements.
Satellite clock error	$\Delta b_i(t)$	Computed using navigation data. There can remain a residual error caused by selective availability for unauthorized users or users without differential GPS connection
User clock bias	b_u	Unknown user clock bias, to be estimated. Once estimated, it may vary only slowly depending on quality of user clock
Satellite position error bias error effect on range	$\Delta D_i(t)$	Unknown bias error cause by errors in GPS control segment estimate

one hour boundaries except for the first data set after a new upload, which may be cut in at any time during the hour. Block II satellite data sets for subframes 1, 2, and 3 are transmitted for periods of two hours before update. Block I satellite subframes 1, 2, and 3 are transmitted for periods of one hour before update.

Words 1 and 2 of each subframe are used for synchronization (preamble), handover word, and C/A code time ambiguity removal. The remaining words, 3–10, of subframe 1 provide clock correction information for the space vehicle clock and space vehicle health and user range accuracy (URA) measures. Subframes 2 and 3 contain ephemeris data that allow estimation of the transmitting satellite's position. Subframes 1, 2, and 3 have the same format from frame to frame, however, subframe 4 and 5 have 25 pages or different sets of data and contain the almanac that gives the approximate satellite ephemeris, clock correction, and space vehicle status for all of the satellites. The almanac data permit the user to select the best set of satellites or simply to determine which satellites are in view. Subframe 4 also contains ionospheric modeling and UTC-GPS clock correction information. A detailed view of the frame structure is shown in Fig. 3. Note that each 30-bit word includes six parity check bits that permit the user receiver to check for errors in the received datastream. A detailed discussion of each of these subframes in Fig. 3 can be found in Sec. II of this chapter, and their mathematical bases are given in the later sections.

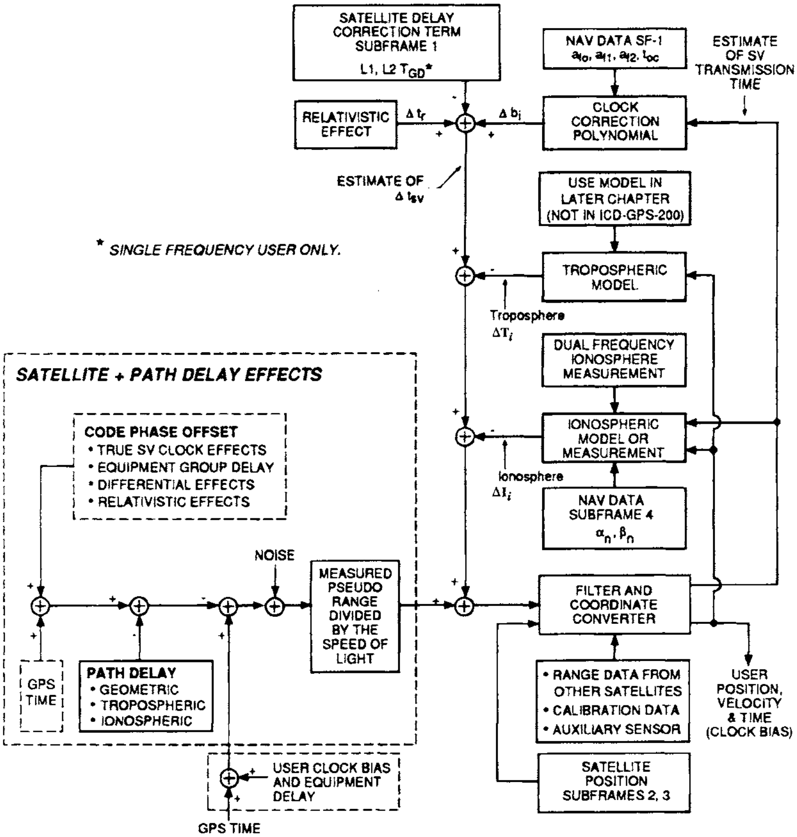

SATELLITE DELAY CORRECTION TERM SUBFRAME 1
L1, L2 T_{GD}*
NAV DATA SF-1
$a_{f0}, a_{f1}, a_{f2}, t_{oc}$
ESTIMATE OF SV TRANSMISSION TIME
RELATIVISTIC EFFECT
Δt_r
Δb_i
CLOCK CORRECTION POLYNOMIAL
ESTIMATE OF Δt_{sv}
USE MODEL IN LATER CHAPTER (NOT IN ICD-GPS-200)
* SINGLE FREQUENCY USER ONLY.
TROPOSPHERIC MODEL
Troposphere ΔT_i
DUAL FREQUENCY IONOSPHERE MEASUREMENT
SATELLITE + PATH DELAY EFFECTS
IONOSPHERIC MODEL OR MEASUREMENT
CODE PHASE OFFSET
• TRUE SV CLOCK EFFECTS
• EQUIPMENT GROUP DELAY
• DIFFERENTIAL EFFECTS
• RELATIVISTIC EFFECTS
Ionosphere ΔI_i
NAV DATA SUBFRAME 4
α_n, β_n
NOISE
MEASURED PSEUDO RANGE DIVIDED BY THE SPEED OF LIGHT
FILTER AND COORDINATE CONVERTER
GPS TIME
PATH DELAY
• GEOMETRIC
• TROPOSPHERIC
• IONOSPHERIC
• RANGE DATA FROM OTHER SATELLITES
• CALIBRATION DATA
• AUXILIARY SENSOR
USER POSITION, VELOCITY & TIME (CLOCK BIAS)
USER CLOCK BIAS AND EQUIPMENT DELAY
SATELLITE POSITION SUBFRAMES 2, 3
GPS TIME

Fig. 1 Navigation data and correction parameters for pseudorange estimate from the pseudorange measurement. This diagram is a modified version of that contained in ICD-GPS-200.[1] The notation *SF* represents the navigation data subframe.

The subframes, frames, and 25-frame superframes are all synchronous with the 1.5-s, *X*1 epochs of the P code. Recall that the full P code has a 1-week period. The superframe also begins at the beginning of each week. Subframes begin at the beginning of the week and are numbered consecutively from the beginning of the week to aid in C/A to P(Y) code acquisition/handover. The timing relationships, shown in Fig. 4 and Fig. 5, illustrate the relationships between *X*1 epochs, time of week (TOW), handover word, *Z*-count, data bits, frames, and the chip durations.

The timing starts at the beginning of each week, defined as midnight Saturday night–Sunday morning in GPS time, which is referenced to UTC time kept by the U.S. Naval Observatory. GPS zero time point is defined as midnight on the night of January 5, 1980/morning of January 6, 1980. Note that the *Z*-count rolls over 1024 weeks later every 19+ years. UTC time is nominally referenced to time at the Greenwich meridian. GPS time differs from UTC in that GPS time does not exhibit the leap second that is sometimes inserted in UTC. GPS time,

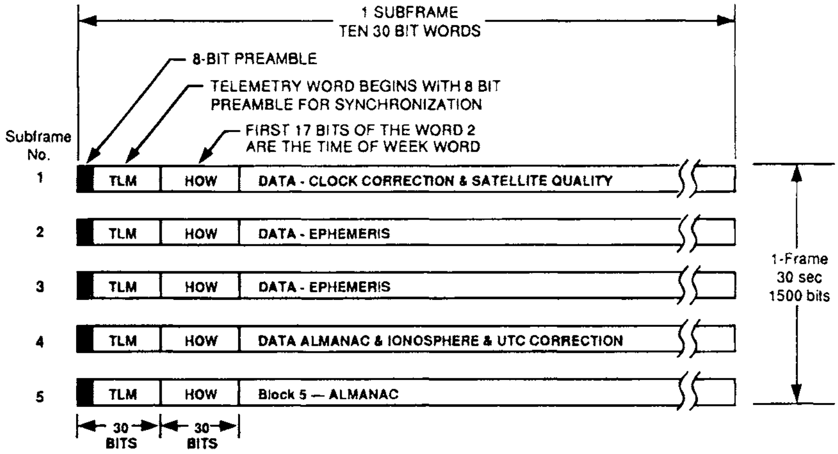

1 SUBFRAME
TEN 30 BIT WORDS
8-BIT PREAMBLE
TELEMETRY WORD BEGINS WITH 8 BIT PREAMBLE FOR SYNCHRONIZATION
FIRST 17 BITS OF THE WORD 2 ARE THE TIME OF WEEK WORD
Subframe No.
1
TLM
HOW
DATA - CLOCK CORRECTION & SATELLITE QUALITY
2
TLM
HOW
DATA - EPHEMERIS
3
TLM
HOW
DATA - EPHEMERIS
4
TLM
HOW
DATA ALMANAC & IONOSPHERE & UTC CORRECTION
5
TLM
HOW
Block 5 — ALMANAC
30 BITS
30 BITS
1-Frame
30 sec
1500 bits

Fig. 2 Simplified GPS frame and subframe format. A superframe consists of 25 frames of 30 s each or 750 s or 12.5 min, and provides a complete almanac.

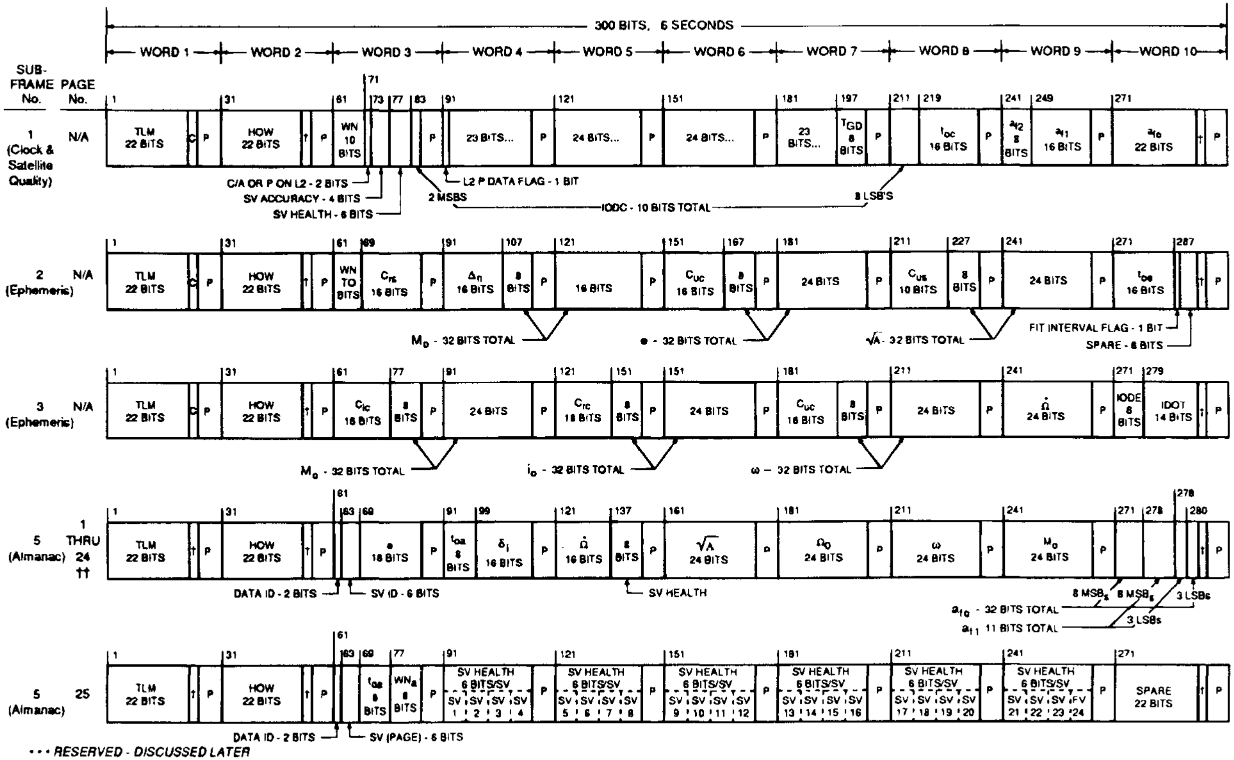
however, is kept to within 1 μs of UTC (modulo-1 s) by the GPS Control Segment. As years pass, the GPS time will differ from UTC by an integer number of seconds.

The number of *X*1 epochs (1.5 s each) since the GPS zero time point modulo-1024 weeks is a 29-bit number called the *Z*-count. The 19 least significant bits of the *Z*-count are referred to as the *time of week* (TOW) count, which is defined as the number of *X*1 epochs (1.5 s each) since the transition from the previous week. A truncated version, the 17 most significant bits of the TOW word is defined as the handover word, which ranges from 0 to 100,799, corresponds to the number of 6-s subframes since the beginning of the week, and is contained in the *L*-band downlink datastream. The HOW removes any timing ambiguity caused by the 1-ms C/A-code period and aids in C/A- to P-code handover/acquisition. The ten most significant digits of the 29-bit *Z*-count represent the number of weeks since the GPS zero time point.

As shown in Fig. 6, each 10-word subframe begins with a telemetry (TLM) word, which in turn begins with an 8-bit preamble, as shown for synchronization (a modified Barker sequence). Other parts of the TLM contain data needed by authorized users, as defined in GPS-ICD-205 and/or GPS-ICD-207.

1. Subframe Synchronization

The 8-bit modified Barker word at the beginning of the TLM word (or the 8-bit modified Barker word plus the two zeros in positions 29 and 30 of the HOW word) provides a synchronization pattern for subframe synchronization by the GPS receiver. However, random data patterns elsewhere in the frame can provide an identical pattern. Furthermore, there is a ±1 sign ambiguity in demodulating any biphase modulated data signal. Thus, the 8-bit or the 8-bit plus 2-bit pattern still has a probability of a "false alarm" or random bit pattern match with the sync pattern of 2×2^{-8} or 2×2^{-10} for 1/128 or 1/512, respectively. Thus,



• • • RESERVED - DISCUSSED LATER

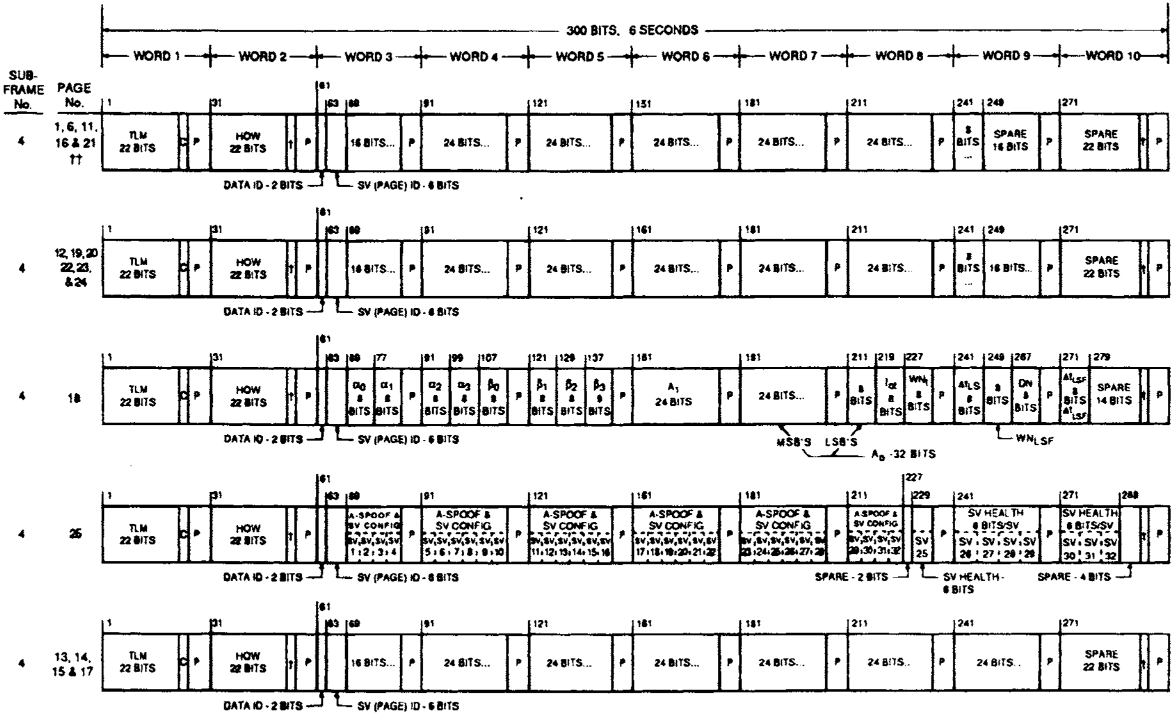
† = TWO NONINFORMATION BEARING BITS USED FOR PARITY COMPUTATION (SEE PARITY CHECK ALGORITHM)

†† = PAGES 2, 3, 4, 5, 7, 8, 9, AND 10 OF SUBFRAME 4 HAVE SAME FORMAT AS PAGES 1 THROUGH 24 OF SUBFRAME 5

P = SIX PARITY BITS

C = TLM BITS 23 AND 24 ARE RESERVED

Figure 3 GPS data format from ICD-GPS-200.[1] Bit 1 is transmitted first. Within each word the most significant bits are transmitted first.



••• RESERVED - DISCUSSED LATER
† = TWO NONINFORMATION BEARING BITS USED FOR PARITY COMPUTATION (SEE PARITY CHECK ALGORITHM)
†† = PAGES 2, 3, 4, 5, 7, 8, 9, AND 10 OF SUBFRAME 4 HAVE SAME FORMAT AS PAGES 1 THROUGH 24 OF SUBFRAME 5
P = SIX PARITY BITS
C = TLM BITS 23 AND 24 WHICH ARE RESERVED

Figure 3 (continued) GPS data format from ICD-GPS-200.[1] Bit 1 is transmitted first. Within each word the most significant bits are transmitted first.

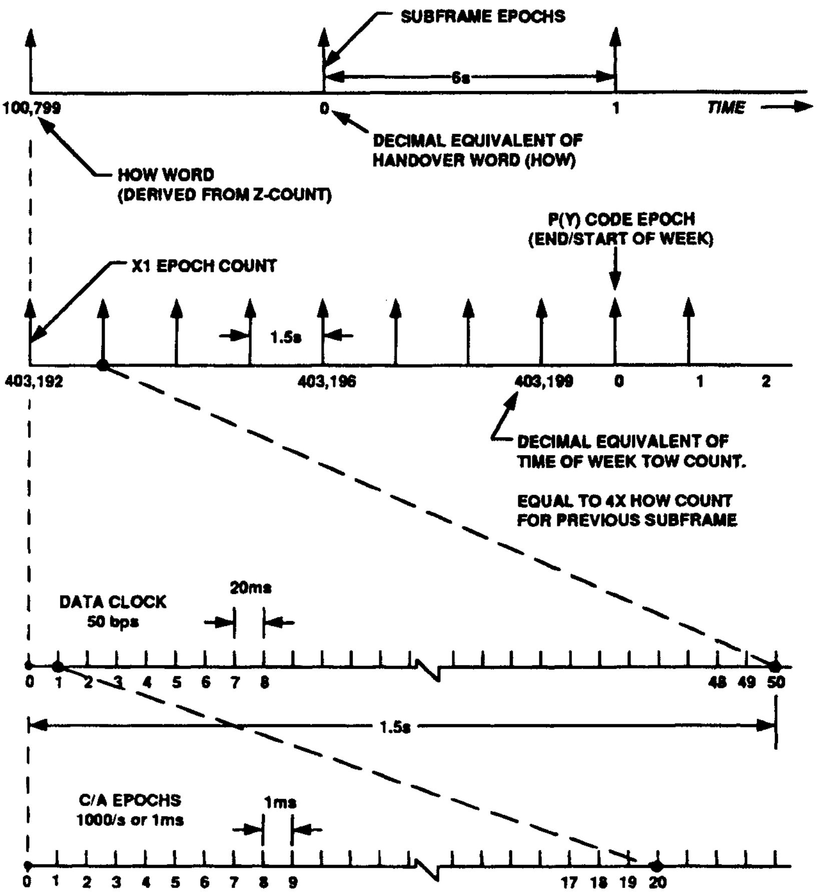

SUBFRAME EPOCHS
6s
100,799
0
1
TIME
DECIMAL EQUIVALENT OF HANDOVER WORD (HOW)
HOW WORD (DERIVED FROM Z-COUNT)
P(Y) CODE EPOCH (END/START OF WEEK)
X1 EPOCH COUNT
1.5s
403,192
403,196
403,199
0
1
2
DECIMAL EQUIVALENT OF TIME OF WEEK TOW COUNT.
EQUAL TO 4X HOW COUNT FOR PREVIOUS SUBFRAME
DATA CLOCK 50 bps
20ms
0 1 2 3 4 5 6 7 8
48 49 50
1.5s
C/A EPOCHS 1000/s or 1ms
1ms
0 1 2 3 4 5 6 7 8 9
17 18 19 20

Fig. 4 Timing relationships between C/A-code epochs, P-code epochs, and navigation data.

the modified Barker word by itself yields too high a false alarm rate to be acceptable. However, we can also check the 17-bit truncated TOW message at the beginning of the HOW word to see that it increments by one and only one from subframe to subframe as a means for confirming the subframe synchronization.

2. *Parity Check Algorithm*

Although the navigation data are normally received at a relatively high signal-to-noise ratio and correspondingly low bit error rates $P_b < 10^{-5}$, it is important to have a parity check algorithm to reject words with any errors in them. Each 30-bit word plus the last two bits of the previous word is encoded into an extended Hamming (32,26) block code of $n = 32$ symbols and $k = 26$ "information" bits, where only 24 of the bits are true information bits d_i, $i \leq 24$.

If the parity transmission bits D_i (computed from the equation below and the $\boldsymbol{H}$ matrix of Fig. 7 or Table 3) do not match the six received parity bits D_i for

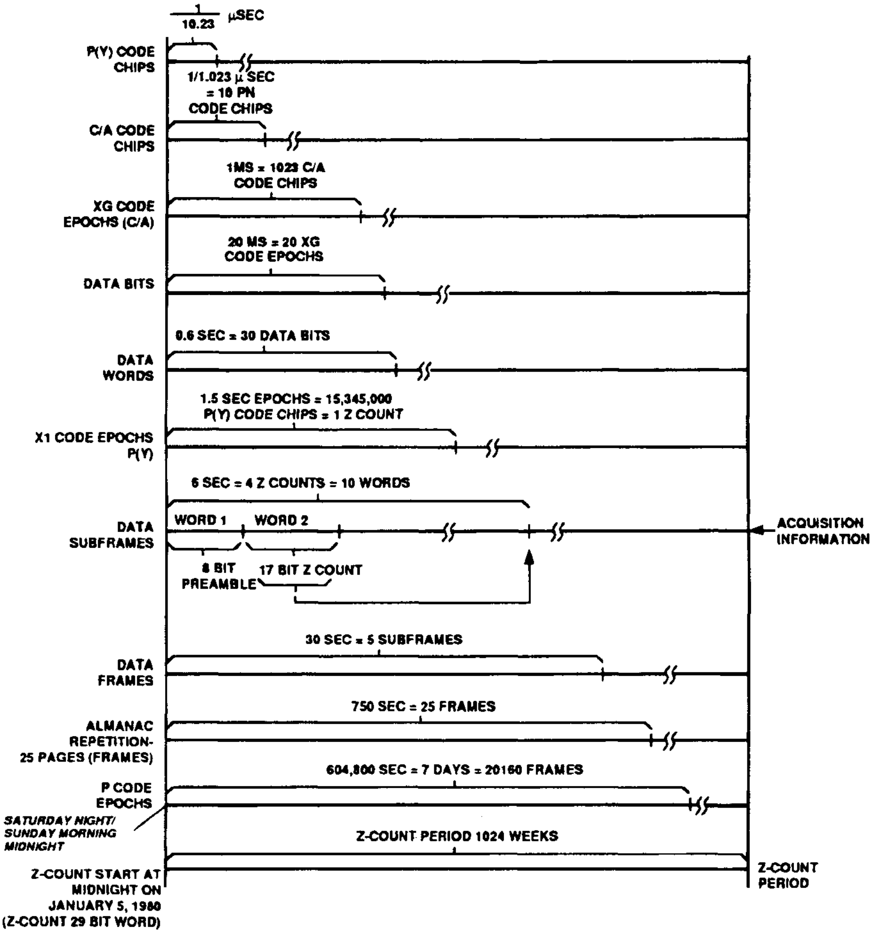

1/10.23 μSEC
P(Y) CODE CHIPS
1/1.023 μ SEC = 10 PN CODE CHIPS
C/A CODE CHIPS
1MS = 1023 C/A CODE CHIPS
XG CODE EPOCHS (C/A)
20 MS = 20 XG CODE EPOCHS
DATA BITS
0.6 SEC = 30 DATA BITS
DATA WORDS
1.5 SEC EPOCHS = 15,345,000 P(Y) CODE CHIPS = 1 Z COUNT
X1 CODE EPOCHS P(Y)
6 SEC = 4 Z COUNTS = 10 WORDS
DATA SUBFRAMES
WORD 1
WORD 2
8 BIT PREAMBLE
17 BIT Z COUNT
ACQUISITION INFORMATION
30 SEC = 5 SUBFRAMES
DATA FRAMES
750 SEC = 25 FRAMES
ALMANAC REPETITION-25 PAGES (FRAMES)
604,800 SEC = 7 DAYS = 20160 FRAMES
P CODE EPOCHS
SATURDAY NIGHT/ SUNDAY MORNING MIDNIGHT
Z-COUNT PERIOD 1024 WEEKS
Z-COUNT START AT MIDNIGHT ON JANUARY 5, 1980 (Z-COUNT 29 BIT WORD)
Z-COUNT PERIOD

Fig. 5 Timing relationships of C/A codes, and data to Z-count.

D_{25}, D_{26}, D_{27}, D_{28}, D_{29}, D_{30}, then the information bits d_i for $i \leq 24$ are rejected. The parity check equation for the six parity check bits is $\boldsymbol{p} = \boldsymbol{Hd}$ where $\boldsymbol{H}$ is the (24×6) matrix of Fig. 7 and $D_i = d_i + D_{30}^*$ for $i \leq 24$, the received databit vector is $\boldsymbol{d} = (D_{29}^*, D_{30}^*, d_1, d_2, \ldots, d_{24})$, and $\boldsymbol{p}$ represents the vector $(D_{25}, D_{26}, D_{27}, D_{28}, D_{29}, D_{30})$.

In making this calculation, the $d_i = D_i \oplus D_{30}^*$, $i \leq 24$ are computed first, then the parity check bits D_i, for $25 \leq i \leq 30$ are computed and checked. In Fig. 7, note that each row in the parity matrix is a cyclic shift of the previous row, except for the last row. The last row is a check of all the previous rows. Note that the sum of all of the row vectors in the matrix is the all "1" row vector.

The Hamming code of length n with parameter m is of the form $(2^m - 1, 2^m - m - 1, 3) = (n, k, 3)$ is of minimum weight 3 for all m. For $m = 5$, the code is (31, 26, 3). The number of information bits in the code word is k. The

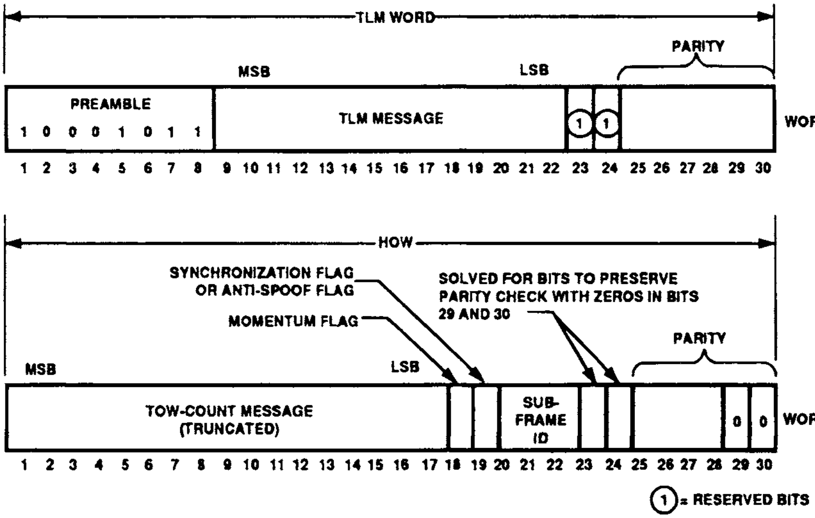


Fig. 6 Telemetry and handover words formats referenced from ICD-GPS-200.[1]

normal Hamming code (31,26) is a perfect code, which means that $2^{n-k} = 1 + n$, that is, $2^{31-26} = 2^5 = 1 + 31$ and has distance 3 for double error detection capability. It takes a minimum of three errors to cause an undetected error.

For GPS the distance 3 Hamming code is converted to a minimum distance 4 code by appending an added parity bit (the 32nd bit) that checks all of the other symbol bits including the other parity bits. In effect, this change adds the last row to the parity check matrix shown in Fig. 7. The code is then shortened to 30 bits by deleting two of the databits.

The GPS parity check code is an extended Hamming code (32,26) and has distance 4. Therefore, it takes certain patterns of four errors to cause an undetectable error.[2-4] If the error probability is p, then the probability of an undetectable error is approximately equal to $p_u \cong 1085\,p^4 - 29295\,p^5 + 403403\,p^6 + 0\,(p^7)$, and it is negligibly small for moderately low error probabilities $p < 10^{-3}$.

$$
\mathbf{H} = \begin{array}{cccccccccccccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \\
\hline
1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1
\end{array}
$$

Fig. 7 Parity matrix *H* for the extended Hamming (32,26) code where $d_1, d_2, \ldots d_{24}$ are the source databits and $D_1, D_2, \ldots D_{30}$ are the bits transmitted by the global positioning system satellite. The notation D_{29}^*, D_{30}^* represents the last 2 bits transmitted in the previous 30-bit word. Note that each row in *H* is simply a cyclic shift of the previous row except for the row D_{30}.

Table 3 Parity encoding equations for each 30-bit word (from GPS-ICD-200)[1]

$D_1 = d_1 \oplus D_{30}{}^*$
$D_2 = d_2 \oplus D_{30}{}^*$
$D_3 = d_3 \oplus D_{30}{}^*$
$\cdot \quad \cdot$
$\cdot \quad \cdot$
$\cdot \quad \cdot$
$D_{24} = d_{24} \oplus D_{30}{}^*$
$D_{25} = D_{29}{}^* \oplus d_1 \oplus d_2 \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_{10} \oplus d_{11} \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{17} \oplus d_{18} \oplus d_{20} \oplus d_{23}$
$D_{26} = D_{30}{}^* \oplus d_2 \oplus d_3 \oplus d_4 \oplus d_6 \oplus d_7 \oplus d_{11} \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{18} \oplus d_{19} \oplus d_{21} \oplus d_{24}$
$D_{27} = D_{29}{}^* \oplus d_1 \oplus d_3 \oplus d_4 \oplus d_5 \oplus d_7 \oplus d_8 \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{19} \oplus d_{20} \oplus d_{22}$
$D_{28} = D_{30}{}^* \oplus d_2 \oplus d_4 \oplus d_5 \oplus d_6 \oplus d_8 \oplus d_9 \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{17} \oplus d_{20} \oplus d_{21} \oplus d_{23}$
$D_{29} = D_{30}{}^* \oplus d_1 \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_7 \oplus d_9 \oplus d_{10} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{17} \oplus d_{18} \oplus d_{21} \oplus d_{22} \oplus d_{24}$
$D_{30} = D_{29}{}^* \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_8 \oplus d_9 \oplus d_{10} \oplus d_{11} \oplus d_{13} \oplus d_{15} \oplus d_{19} \oplus d_{22} \oplus d_{23} \oplus d_{24}$

where $d_1, d_2, \ldots d_{24}$, are the source data bits;
the symbol (*) is used to identify the last 2 bits of the previous word of the subframe;
$D_{25}, \ldots D_{30}$ are the computed parity bits;
$D_1, D_2, D_3, \ldots D_{29}, D_{30}$, are the bits transmitted by the space vehicle (SV); and $\oplus$ is the modulo-2 or "exclusive-or" operation.

II. Detailed Description of the Navigation Data Subframes

The previous section defined the overall format of the 50-bs navigation data. This section describes in detail each of the elements of the navigation data for each of the five subframes. The next section provides some of the analytical background for the satellite clock and ephemeris calculations.

A. Subframe 1—GPS Clock Correction and Space Vehicle Accuracy Measure

Subframe 1 contains the data to be used in the`algorithms described below to provide the space vehicle clock correction. It also contains data to give an estimate of the effect of space vehicle accuracy on user range accuracy (URA).

1. GPS Clock Correction Data Formats—Subframe 1

The user receiver needs to correct the GPS satellite clock errors. The user receiver must have an accurate representation of GPS system time t at the time of transmission for the GPS signal it now is receiving from satellite i. The satellite clock correction Δt_{sv} is obtained using coefficients broadcast from the satellite after being uploaded by the GPS control segment. The control segment actually uploads several different sets of coefficients to the satellite, of which each set is valid over a given time period. The data sets are then transmitted in the downlink datastream to the users in the appropriate time intervals. Subframe 1, words 8, 9, 10, shown previously in Fig. 3, contain the data needed by the users to perform corrections of the space vehicle clock. These corrections represent a second-order polynomial in time. Specifically, bits 9–24 of word 8, bits 1–24 of word

9, and bits 1–22 of word 10 provide four clock correction parameters, t_{oc}, a_{f2}, a_{f1}, a_{f0}, which are described in the following paragraphs.

The GPS time t (the space vehicle SV clock time) needed to solve for user position is $t = t_{SV} - \Delta t_{SV}$, where t_{SV} is the SV pseudorandom noise (PRN) code phase time at the time of transmission and is easily determined by the GPS receiver. The satellite clock correction term is approximated by a polynomial $\Delta t_{SV} = a_{f0} + a_{f1}(t - t_{oc}) + a_{f2}(t - t_{oc})^2 + \Delta t_R$, where a_{f0}, a_{f1}, and a_{f2} are the polynomial correction coefficients corresponding to phase error, frequency error, and rate of change of frequency error; the relativistic correction is Δt_R; and t_{oc} is a reference time (in s) for clock correction. Table 4 describes the parameters in number of bits, scale factors, and units.

The relativistic correction must be computed by the user. A first-order effect described in the GPS ICD[1] gives the relativistic correction for an Earth-centered, Earth-fixed (ECEF) observer and a GPS satellite of orbit eccentricity e. This relativistic correction varies as the sine of the satellite eccentric anomaly E_k as follows:

$$\Delta t_R = Fe\sqrt{A} \sin E_k = 2\boldsymbol{R} \cdot \boldsymbol{V}/c^2$$

where

F = $-2\sqrt{\mu}/c^2 = -4.442807633 \times 10^{-10}\ s/m^{1/2}$
μ = 3.986005×10^{14} m^3/s^2 value of Earth universal gravitational parameter
c = 2.99792458×10^8 m/s
$\boldsymbol{R}$ = instantaneous position vector of the space vehicle
$\boldsymbol{V}$ = instantaneous velocity vector of the space vehicle
e = space vehicle orbit eccentricity

Table 4 Subframe 1 parameters for clock correction and other data

Parameter	No. of bits	Scale factor, LSB[a]	Effective range[b]	Units
Code on L_2	2	1	——	N/A
Week no.	10	1	——	Week
L_2 P data flag	1	1	——	Discretes
SV accuracy	4	——	——	(see text)
SV health	6	1	——	N/A
T_{GD}	8[c]	2^{-31}	——	s
Issue of data clock (IODC)	10	——	——	(see text)
t_{OC}	16	2^4	604,784	s
a_{f2}	8[c]	2^{-55}	——	s/s^2
a_{f1}	16[c]	2^{-43}	——	s/s
a_{f0}	22[c]	2^{-31}	——	s

[a] Least significant bits.
[b] Unless otherwise indicated in this column, effective range is the maximum range attainable with indicated bit allocation and scale factor.
[c] Parameters so indicated shall be two's complement, with the sign bit (+ or −) occupying the most significant bits (MSB).

E_k = eccentric anomaly of the satellite orbit
A = semimajor axis of the satellite orbit

R and V are expressed in the same inertial coordinate system. Chapter 18, this volume, discusses the relativistic effects and the derivation of this equation in more detail.

In addition, as discussed in Chapter 18, the other relativistic effects are as follows:

1) Increase in the received clock frequency by a fixed user on the surface of the Earth's geoid by a fraction $\Delta f/f = +4.46 \times 10^{-10}$. This effect is compensated by purposely setting the 10.23 MHz satelite clock low by $\Delta f = +4.56 \times 10^{-3}$ Hz. Thus, the satellite clock is set to 10,229,999.99543 Hz before launch to ensure that the received GPS signals arrive at the Earth geoid at the correct frequencies.

2) We may also have to account for any significant velocity of the user relative to the Earth or displacement in altitude (gravitational potential) from the surface of the geoid. For example, if the user is at an altitude above the geoid, the fractional increase in the received satellite frequency is not as large. Some of these effects simply may be accounted for by a modification in the user clock bias offset and may not significantly affect user position estimates, because they are approximately the same for all satellites.

We must also account for the rotation of the Earth during the time of transit of the GPS signals from satellite to user. It has already been pointed out that the satellite position has been computed at the time of transmission, whereas the user receiver is computing position at a slightly later time. The Earth has rotated during this transit time, and this rotation must be taken into account. Note that these times of transit for different satellites are not all identical. This effect is a simple effect of the finite velocity of light.

2. *$L_1 - L_2$ Correction—Single-Frequency Users*

The $L_1 - L_2$ delay correction term is calculated by the GPS Control Segment (CS) to account for the group delay difference in the space vehicle transmission between L_1 and L_2 signals based on measurements made on the SV prior to launch in the factory test. The GPS CS uses a two-frequency ionospheric correction and estimates the a_{f_0} satellite clock correction term based on the dual-frequency measurement. Thus, the user employing both L_1 and L_2 in the ionospheric correction need make no further correction. However, the user who employs only L_1 must modify the space vehicle clock correction by $(\Delta t_{sv})_{L_1} = \Delta t_{sv} - T_{GD}$, where T_{GD} is provided in the subframe 1 data by bits 17–24 of word 7 (see Table 4). For the user who employs only L_2 the space vehicle clock correction is $(\Delta t_{sv})_{L_2} = \Delta t_{sv} - \Gamma T_{GD}$ where $\Gamma = (f_{L_1}/f_{L_2})^2 = (1575.42/1227.6)^2 = (77/60)^2$.

The value of the correction term T_{GD} is not equal to the SV group delay differential but rather $T_{GD} = (t_{L_1} - t_{L_2})/(1 - \Gamma)$, where $t_{L_1} - t_{L_2}$ is the SV differential group delay for the satellite. Thus, the value of T_{GD} is not equal to the mean SV group delay differential but to that delay multiplied by $1/(1 - \Gamma)$.

3. *Subframe 1—Space Vehicle Accuracy—User Range Accuracy Index*

Bits 13 through 16 of word 3 subframe 1 give the user range accuracy (URA) index, of the SV for the user who does not have access to the full accuracy of

Table 5 Table of user range accuracy index *N*, vs the user range accuracy interval in meters

URA index	URA, m
0	0.0–2.4
1	2.4–3.4
2	3.4–4.85
3	4.85–6.85
4	6.85–9.65
5	9.65–13.65
6	13.65–24.0
7	24.0–48.0
8	48.0–96.0
9	96.0–192.0
10	192.0–384.0
11	384.0–768.0
12	768.0–1536.0
13	1536.0–3072.0
14	3072.0–6144.0
15	>6144.0[a]

[a](No accuracy prediction is available. Unauthorized users are advised to use the SV at their own risk.)

Table 6 Ephemeris data definitions[1]

Symbol	Definition
M_0	Mean anomaly at reference time
Δ_n	Mean motion difference from computed value
e	Eccentricity of the orbit
$(A)^{1/2}$	Square root of the semimajor axis
$(\mathrm{OMEGA})_0$	Longitude of ascending node of orbit plane at reference time
I_0	Inclination angle at reference time
ω	Argument of perigee
OMEGADOT	Rate of right ascension
IDOT	Rate of inclination angle
C_{uc}	Amplitude of the cosine harmonic correction term to the argument of latitude
C_{us}	Amplitude of the sine harmonic correction term to the argument of latitude
C_{rc}	Amplitude of the cosine harmonic correction term to the orbit radius
C_{rs}	Amplitude of the sine harmonic correction term to the orbit radius
C_{ic}	Amplitude of the cosine harmonic correction term to the angle of inclination
C_{is}	Amplitude of the sine harmonic correction term to the angle of inclination
T_{oe}	Reference time for ephemeris
IODE	Issue of data (ephemeris)

GPS. (These nonmilitary users are termed the unauthorized users in ICD-GPS-200.[1]) The URA itself (as opposed to the index) is given in meters. The URA index *N* is an integer in the range of 0–15 and has the relationship to the URA of the SV shown in Table 5.

4. *Issue of Data-Clock*

The issue of data clock (IODC) indicates the issue number of the data set for clock correction, which provides a means for detecting any change in the clock correction parameters. This information is carried in bits 23 and 24 of word 3 MSB and bits 1–8 of word 8 in subframe 1.

B. GPS Ephemeris Parameters—Subframes 2 and 3

The purely elliptical Kepler orbit is precise only for a simple two-body problem where the mutual gravitational attraction between the two bodies is the only force involved. In the actual GPS satellite orbit, there are many perturbations to the ideal orbit, including nonspherical Earth gravitational harmonics; lunar, solar gravitational attraction; and solar flux. Thus, the GPS orbit is modeled as a modified elliptical orbit with correction terms to account for these perturbations: 1) sin, cos perturbations to the a) argument of latitude, b) orbit radius, and c) angle of inclination; and 2) rate of change of a) right ascension, and b) inclination angle.

Furthermore, the parameters for this model are changed periodically to give a best fit to the actual satellite orbit. In normal operations, the fit interval is 4 hours. Subframes 2 and 3 provide 375 bits of information for the modified Keplerian model. Table 6 shows ephemeris model parameters including the sinusoidal perturbations to the orbit radius, the angle of inclination and argument of latitude; the rate of change of inclination angle, angular rate of change of the right ascension; and the basic Keplerian parameters. The scale factors for these parameters are given in Table 7.

1. *Calculation of Satellite Position*

By demodulating and extracting the navigation data in subframes 2 and 3, the user can calculate the satellite position vs. time. The issue of data-ephemeris IODE is a number provided in both subframes 2 and 3 for purposes of comparison and for comparison with the 8 LSB of the IODC term in subframe 1. It should be pointed out that the two IODE numbers in subframes 2 and 3 must match and should also correspond to the IODC for the clock in subframe 1; otherwise, a data set cutover has occurred, and the user must collect new data.

2. *Curve Fit Intervals for the Ephemeris Data*

Bit 17 in word 10 of subframe 2 is a fit interval flag that indicates whether the GPS CS used a least squares fit over a 4-h period or a longer 6-h period; a “0” bit is transmitted for fit periods greater than 4 h. For data sets with a 4-h fit interval (transmitted during the first approximately 1-day period after upload), the curve fit procedures provide a URE contribution for the predicted SV ephemeris of

Table 7 Ephemeris parameters[1]

Parameter	Number of bits[a]	Scale factor, LSB	Effective range[b]	Units
IODE	8	——	——	(See text)
C_{rs}	16[c]	2^{-5}	——	m
Δn	16[c]	2^{-43}	——	Semicircles
M_0	32[c]	2^{-31}	——	Semicircles
C_{uc}	16[c]	2^{-29}	——	rad
e	32	2^{-33}	0.03	Dimensionless
C_{us}	16[c]	2^{-29}	——	rad
$(A)^{1/2}$	32	2^{-19}	——	$m^{1/2}$
t_{oe}	16	2^{4}	604,784	s
C_{ic}	16[c]	2^{-29}	——	rad
$(\text{OMEGA})_0$	32[c]	2^{-31}	——	Semicircles
C_{is}	16[c]	2^{-29}	——	rad
i_0	32[c]	2^{-31}	——	Semicircles
C_{rc}	16[c]	2^{-5}	——	m
ω	32[c]	2^{-31}	——	Semicircles
OMEGADOT	24[c]	2^{-43}	——	Semicircles/s
IDOT	14[c]	2^{-43}	——	Semicircles/s

[a] See Fig. 3 for complete bit allocation in subframe.
[b] Unless otherwise indicated in this column, effective range is the maximum range attainable with indicated bit allocation and scale factor.
[c] Parameters so indicated shall be two's complement, with the sign bit (+ or −) occupying the MSB.

less than 0.35 m, one sigma. These URE component values apply when the data set is transmitted, as well as for a period of 3 h thereafter. The longer, less accurate, 6-h fit interval normally is not used. It is employed if the upload does not occur daily, and the same uploaded data set must apply for a 2nd day through the 14th day after upload. For data sets with a 6-h fit interval, the curve fit provides a URE of less than 1.5 m, one sigma. These URE values apply during transmission and for 2 h thereafter.

The equations in Table 8 give the space vehicle antenna phase center position in WGS-84[5] Earth-centered, Earth-fixed reference frame (including correction for the Earth's rotation with the x_k' to x_k matrix transformation). The ECEF coordinate system is defined as follows:

Origin = Earth center of mass (Geometric center of the WGS-84[5] ellipsoid)

z axis = Parallel to the direction of the Conventional International Origin (CIO) for polar motion as defined by the Bureau International de l'Heure (BIH) on the basis of the latitudes adopted for the BIH stations (Rotation axis of the WGS-84[5] ellipsoid)

x axis = Intersection of the WGS-84[5] reference meridian plane and the plane of the mean astronomic equator, the reference meridian being parallel to the zero meridian defined by the Bureau International de l'Heure on the basis of the longitudes adopted for the BIH stations

y axis = Completes a right-handed, Earth-centered, Earth-fixed orthogonal coordinate system measured in the plane of the mean astronomic equator 90° east of the x axis (x, y axis of the WGS-84[5] ellipsoid)

Table 8 Elements of ephemeris model equations[1]

$\mu = 3.986005 \times 10^{14}\ \mathrm{m^3/s^2}$	WGS-84[5] value of the Earth's universal gravitational parameter	
$\dot{\Omega}_e = 7.2921151467 \times 10^{-5}\ \mathrm{rad/s}$	WGS-84[5] value of the Earth's rotation rate	
$A = (\sqrt{A})^2$	Semimajor axis	
$n_0 = \sqrt{\mu/A^3}$	Computed mean motion–rad/s	
$t_k = t - t_{oe}{}^{a}$	Time from ephemeris reference epoch	
$n = n_0 + \Delta n$	Corrected mean motion	
$M_k = M_o + nt_k$	Mean anomaly	
$\pi = 3.1415926535898$	GPS standard value for π	
$M_k = E_k - e \sin E_k$	Kepler's equation for the eccentric anomaly E_k (may be solved by iteration), rad	
$\nu_k = \tan^{-1}\left\{\dfrac{\sin \nu_k}{\cos \nu_k}\right\} = \tan^{-1}\left\{\dfrac{\sqrt{1-e^2}\sin E_k/(1 - e\cos E_k)}{(\cos E_k - e)/(1 - e \cos E_k)}\right\}$	True anomaly υ_k as a function of the eccentric anomaly	
$E_k = \cos^{-1}\left\{\dfrac{e + \cos \nu_k}{1 + e \cos \nu_k}\right\}$	Eccentric anomaly	
$\Phi_k = \nu_k + \omega$	Argument of latitude	
$\delta u_k = C_{us} \sin 2\Phi_k + C_{uc} \cos 2\Phi_k$	Argument of latitude correction	Second harmonic perturbations
$\delta r_k = C_{rs} \sin 2\Phi_k + C_{rc} \cos 2\Phi_k$	Radius correction	
$\delta i_k = C_{is} \sin 2\Phi_k + C_{ic} \cos 2\Phi_k$	Inclination correction	
$u_k = \Phi_k + \delta u_k$	Corrected argument of latitude	
$r_k = A(1 - e \cos E_k) + \delta r_k$	Corrected radius	
$i_k = i_0 + \delta i_k + (\mathrm{IDOT})\, t_k$	Corrected inclination	
$x_k' = r_k \cos u_k$ $y_k' = r_k \sin u_k$	Satellite position in orbital plane	
$\Omega_k = \Omega_0 + (\dot{\Omega} - \dot{\Omega}_e)\, t_k - \dot{\Omega}_e\, t_{oe}$	Corrected longitude of ascending node	
$x_k = x_k' \cos \Omega_k - y_k' \cos i_k \sin \Omega_k$ $y_k = y_k' \sin \Omega_k + y_k' \cos i_k \cos \Omega_k$ $z_k = y_k' \sin i_k$	Satellite position in Earth-centered, Earth-fixed coordinates	

[a] t is GPS system time at time of transmission; i.e., GPS time corrected for transit time (range/speed of light). Furthermore, t_k shall be the actual total time difference between the time t and the epoch time t_{oe} and must account for beginning or end of week crossovers. That is, if t_k is greater than 302,400 s, subtract 604,800 s from t_k. If t_k is less than 302,400 s, add 604,800 s to t_k.

From Table 8 note that it is the mean anomaly $M_k = M_0 + nt_k$ that varies linearly with the time interval t_k. However, the solution of the satellite position requires knowledge of the eccentric anomaly E_k, which does not vary linearly with time unless the eccentricity $e = 0$.

The eccentric anomaly E_k must be solved for by iterative calculations not given in Table 8. However, some of these techniques are briefly discussed in Sec. III of this chapter. Note again, that the model of Table 8 is not merely a simple elliptical orbit. Second harmonic (sinusoidal) corrections are made for the argument of latitude, radius, and inclination of the Kepler model for satellite position. These corrections are then introduced to provide the corrected position for the satellite in the orbital plane, the corrected radius, argument of latitude, and inclination. Finally, the *x, y, z* coordinates for the satellite position are transferred to Earth-centered, Earth-fixed coordinates to be used in the final computations of the user position.

3. Geometric Range

The ICD-GPS-200[1] also states "The user must also account for the effects due to Earth rotation during the time of signal propagation so as to evaluate the path delay in an inertially stable system." This effect is discussed later in Chapter 18 on relativistic effects.

C. Subframes 4 and 5—Almanac, Space Vehicle Health, and Ionosphere Models

1. Almanac Data

Almanac data are used for satellite selection purposes and as aids to acquisition; the almanac can also be used to give approximate Doppler and delay information. Almanac data are used by P(Y) code users in order to perform direct P(Y) code acquisition (if they choose not to acquire the C/A signal first and then handover to the P(Y) code). The almanac data provides a truncated, reduced precision set of the ephemeris parameters described earlier in Table 6. Almanac data provide approximate ephemeris information for up to 32 satellites along with the associated health data for each satellite. Almanacs are provided only for valid satellites or perhaps for a satellite that is about to become active. Where there is no satellite data to fill an almanac data slot, dummy alternating "0s" and "1s" are transmitted to aid in synchronization.

Subframes 4 and 5 each carry 25 pages of information, one new page per frame repetition. We term a 25-frame segment a superframe. Thus, a GPS receiver must demodulate 25 frames over a period of 25 × 30 s or 12.5 min in order to receive all 25 pages of the subframe 4 and 5 almanac data. Of particular interest are the pages shown in Table 9.

Table 9 Key elements of pages in subframes 4 and 5

Pages	Subframe 4	Pages	Subframe 5
2,3,4,5,7,8,9,10	Almanac data for SV 25 through 32	1–24	Almanac data for SV 1–24
18	Ionosphere and UTC data	25	SV health for SV 1–24, almanac reference time and reference week number
25	Anti-spoof flag SV configuration for 32 SV, SV health for SV 25–32	——	——
Other pages	Reserved + special messages, spares	——	——

Table 10a Almanac parameters and accuracy

Age of data time since transmission	Almanac accuracy
1 day	900 m
1 week	1200 m
2 weeks	3600 m

The almanac data are much less accurate than the detailed ephemeris data of subframes 2 and 3. However, the almanac data are valid for longer periods of time and do not require frequent updates. Approximate one sigma almanac accuracy varies as a function of the time since the time of transmission approximately as shown in Table 10a.

The almanac parameters and their scale factors are shown in Table 10b. The algorithm for these parameters is the same ephemeris algorithm as discussed for subframes 2 and 3. Where the almanac does not include a parameter; e.g., sinusoidal corrections, these parameters are set to zero. For the inclination angle, a nominal value of 0.30 semicircles is implicit, and only a parameter δ_i, the correction to the inclination, is transmitted.

In addition, the almanac provides truncated clock correction a_{f0}, a_{f1} parameters for the algorithm discussed in subframe 1. The Almanac time correction provides time to within 2 μs of GPS time using the first order polynomial $t = t_{SV} - \Delta t_{SV}$ where t is GPS time, t_{SV} is the space vehicle clock time (PRN code phase at message transmission time) and $\Delta t_{SV} = a_{f0} + a_{f1}\, t_k$, where t_k is the time from epoch.

The almanac data occupy almost all bits of words 3–10 of each page of subframe 5 (pages 1–24) and subframe 4 (pages 2–5 and 7–10). The exceptions are the first 8 bits of word 3 (data ID and SVID), bits 17–24 of word 5 (SV health), and the 50 bits of parity. The "0" SVID, binary all zeroes, is used to

Table 10b Almanac parameters

Parameter	Number of bits	Scale factor LSB	Effective range	Units
ϵ	16	2^{-21}		dimensionless
t_{oa}	8	2^{-12}	602,112	s
δ_i	16[a]	2^{-19}	——	semicircles
OMEGADOT	16[a]	2^{-38}	——	semicircles/s
$(A)^{1/2}$	24	2^{-11}	——	$m^{1/2}$
(OMEGA)0	24[a]	2^{-11}	——	semicircles
ω	24[a]	2^{-23}	——	semicircles
M_0	24[a]	2^{-23}	——	semicircles
a_{f0}	11[a]	2^{-20}	——	s
a_{f1}	11[a]	2^{-38}	——	s/s

[a] Parameters so indicated shall be two's complement, with the sign bit (+ or −) occupying the MSB.

identify a dummy satellite. When not all satellite slots are needed for different satellites, the same satellite almanac data may be repeated in more than one page. Space vehicle ID is given by bits 3–8 of word 3.

3. *Space Vehicle Health*

Subframes 4 and 5 also contain two types of space vehicle health data: 1) each of the 32 pages that contain satellite clock/ephemeris data also provide an 8-bit SV health status for that particular satellite; and 2) the 25th page of subframe 4 and 5 jointly contain a satellite summary consisting of 6-bit health status words for up to 32 space vehicles.

The first three most significant bits of the 8-bit health words give the health of the navigation data for that space vehicle; e.g., an indication that the *Z*-count in the HOW word is good or bad. The five LSB of the 8-bit words and the 6-bit words give the health of the space vehicle signal components, as described in Table 11.

4. *Translation of GPS Time to UTC Time*

GPS time is based on atomic standard time, and the time broadcast from the satellite is continuous (modulo-1 week) without the leap seconds of UTC, because the introduction of leap seconds would throw the P-code receivers out of lock at the time when they are introduced. Nonetheless, GPS time is maintained by the GPS CS to be within 1 μs of UTC (USNO) time (modulo-1 s) and provides correction parameters in the GPS navigation message. Thus, the GPS provides an important time transfer function. The UTC–GPS translation parameters are shown in Table 12.

The correction parameters to convert GPS time broadcast by the satellite to UTC are contained in the 24 MSB of words 6–9 plus the 8 MSB of word 10 in page 18 of subframe 4. The bit length scale factors are shown in Table 12.

The information contains the parameters required to relate GPS time to UTC and provides notice to the user of any delta time in the recent past or the near future due to leap seconds Δt_{LS} and the week number $\mathrm{WN}_{\mathrm{LSF}}$ at which that leap second becomes effective. The above relationships apply for the vast majority of the time. However, when the user is operating at a time near the time for a leap second change, special adjustments are required.

The algorithm defining the relationship between GPS time and UTC using the navigation data in subframe 4 is as follows:[1]

$t_{\mathrm{UTC}} \quad = (t_E - \Delta t_{\mathrm{UTC}})$ [modulo-86,400 s]

where t_{UTC} is in ss and

$\Delta t_{\mathrm{UTC}} \quad = \Delta t_{\mathrm{LS}} + A_0 + A_1(t_E - t_{\mathrm{ot}} + 604{,}800\,(\mathrm{WN} - \mathrm{WN}_t))$, s

t_{E} = GPS time as estimated by the user on the basis of correcting t_{SV} for factors given in the Subframe 1 clock correction discussion as well as for ionospheric and SA (dither) effects

Δt_{LS} = delta time due to leap seconds

A_0 and A_1 = constant and first-order terms of polynomial

t_{ot} = reference time for UTC data (see Table 13)

Table 11 Codes for health of space vehicle signal components

MSB				LSB	Definition
0	0	0	0	0	All signals OK
0	0	0	0	1	All signals weak[a]
0	0	0	1	0	All signals dead
0	0	0	1	1	All signals have no data modulation
0	0	1	1	0	L_1 P signal weak
0	0	1	0	1	L_1 P signal dead
0	0	1	1	0	L_1 P signal has no data modulation
0	0	1	1	1	L_2 P signal weak
0	1	0	0	0	L_2 P signal dead
0	1	0	0	1	L_2 P signal has no data modulation
0	1	0	1	0	L_1 C signal weak
0	1	0	1	1	L_1 C signal dead
0	1	1	0	0	L_1 C signal has no data modulation
0	1	1	0	1	L_2 C signal weak
0	1	1	1	0	L_2 C signal dead
0	1	1	1	1	L_2 C signal has no data modulation
1	0	0	0	0	L_1 and L_2 P signal weak
1	0	0	0	1	L_1 and L_2 P signal dead
1	0	0	1	0	L_1 and L_2 P signal has no data modulation
1	0	0	1	1	L_1 and L_2 C signal weak
1	0	1	0	0	L_1 and L_2 C signal dead
1	0	1	0	1	L_1 and L_2 C signal has no data modulation
1	0	1	1	0	L_1 signal weak[a]
1	0	1	1	1	L_1 signal dead
1	1	0	0	0	L_1 signal has no data modulation
1	1	0	0	1	L_2 signal weak[a]
1	1	0	1	0	L_2 signal dead
1	1	0	1	1	L_2 signal has no data modulation
1	1	1	0	0	SV *Is* temporarily out (do not use this SV during current pass[a])
1	1	1	0	1	SV *Will Be* temporarily out (use with caution[a])
1	1	1	1	0	Spare
1	1	1	1	1	More than one combination would be required to describe anomalies (except those marked by[a])

[a] Three to six-dB below specified power level due to reduced power output, excess phase noise, SV attitude, etc.

WN = current week number (derived from subframe 1)
WN_t = UTC reference week number

Note that the number of seconds in a day is 86,400.

The estimated GPS time is in seconds relative to end/start of week. The reference time for UTC data t_{ot} is referenced to the start of that week whose number is given in word 8 of page 18 in subframe 4, which represents the 8 LSB of the week number. The user must account for the truncated nature of the week number (see ICD-GPS-200[1]).

When the effectivity time of the leap second event is in the past relative to the user's current time, the relationship presented above is valid except that Δt_{LSF}

Table 12 GPS-UTC clock correction parameters from subframe 4, page 18[a]

Parameter	Number of bits	Scale factor LSB	Effective range	Units
A_0	32[b]	2^{-30}	——	s
A_1	24[b]	2^{-50}	——	s/s
Δt_{LS}	8[b]	1	——	s
t_{ot}	8	2^{12}	602,112	s
WN_t	8	1	——	weeks
WN_{LSF}	8	1	——	weeks
DN	8[c]	1	7	days
$\Delta \mathrm{t}_{\mathrm{LSF}}$	8[b]	1	——	s

[a] The notations DN and WN stand for day number and week number, respectively.
[b] Parameters so indicated shall be two's complement, with the sign bit (+ or −) occupying the MSB.
[c] Right justified.

Table 13 Reference times for block II satellite vehicles to be used for various clock, ephemeris, almanac, and UTC correction polynomials.[a]

		Hours after first valid transmission time			
Fit interval, h	Transmission interval, h	t_{oc} clock	t_{oe} ephemeris	t_{oa} almanac	t_{ot} UTC
4	2	2	2	——	——
6	4	3	3	——	——
8	6	4	4	——	——
14	12	7	7	——	——
26	24	13	13	——	——
50	48	25	25	——	——
74	72	37	37	——	——
98	96	49	49	——	——
122	120	61	61	——	——
146	144	73	73	——	——
——	——	——	——	——	——
144 (6 days)	144	——	——	84	84
144 (6 days)	4080	——	——	84	84

[a] This table describes the nominal selection that is expressed modulo-604,800 s in the navigation message.

is substituted for Δt_{LS}. The exception to the above algorithm occurs whenever the user's current time falls within the timespan of $DN + 3/4$ to $DN + 5/4$ where DN is the day number. In this time interval, proper accommodation of the leap second event with a possible week number transition is provided by the following expression for UTC: $t_{\mathrm{UTC}} = W[\text{modulo-}(86400 + \Delta t_{\mathrm{LSF}} - \Delta t_{\mathrm{LS}})]$, s, where $W = (t_E - \Delta t_{\mathrm{UTC}} - 43200)$ [modulo-86400] + 43200, s; and the definition of Δt_{UTC} (as given in the paragraph above) applies throughout the transition period. Note that when a leap second is added, unconventional time values of the form 23:59:60.xxx are encountered. Some user equipment may be designed to approximate UTC by decrementing the running count of time within several seconds after the event, thereby promptly returning to a proper time indication. Whenever a leap second event is encountered, the user equipment must consistently implement carries or borrows into any year/week/day counts.

5. *Subframe 4 Ionospheric Delay Corrections*

Ionospheric group delay can cause a significant error in the measured pseudorange group delay by perhaps as much as 300 ns during the daytime at low elevation angles and a lesser error, perhaps 5–15 ns, delay at nighttime and varies in roughly a diurnal pattern. Ionospheric delay effects are discussed in detail in Chapter 12, this volume. The two-frequency user with access to the P(Y) code can correct for most of this delay error by measurement.

The single-frequency user has several alternatives:

1) Ignore the ionosphere and accept the ionospheric error,

2) Use a model of the ionosphere,

3) Use a single-frequency carrier/code differential ionospheric measurement scheme,

4) Use a dual-frequency codeless technique either by cross-correlating L_1 and L_2 channels or by squaring the L_2 P(Y) code to recover the pure carrier as shown at the end of this chapter. (Carrier frequency estimates from L_1 and the almanac can be used to reject the multiple access interference caused by other carriers in the estimation process.)

The single-frequency user can get an approximate correction by using a model of the ionosphere with model parameters transmitted in the downlink datastream. However, the user should be advised that the ionosphere varies in a manner that is difficult to predict, hence the model provides only an approximate correction (for perhaps 70% of the ionosphere delay). In addition, the single-frequency user must correct for the L_1–L_2 delay differential T_{GD} in the satellite that is not needed by the two-frequency user.

The ionospheric group delay model developed for GPS by Klobuchar (see Chapter 12, this volume) essentially employs a half cosine approximation, as shown in Fig. 8. The ionospheric group delay is modeled essentially as follows:

$$\begin{aligned} I(t) &= F(E)\left[5 \times 10^{-9} + \mathrm{AMP(L)}\cos[2\pi(t - 50{,}400)/\mathrm{PER(L)}] \text{ for day}\right] \\ &= F(E)(5 \times 10^{-9}) \text{ for night} \end{aligned} \tag{3}$$

where $F(E)$ is the obliquity factor that gives a larger I for lower elevation angles E. The parameter AMP(L) is the amplitude of the half cosine for daytime, which is a function of the geomagnetic latitude L of the Earth projection of the ionosphere contact point and PER(L) is the period of the half cosine.

The small constant level of delay equal to 7 ns in this example of Fig. 8 is meant to represent the delay at night. As the sun rises and sets, the ionospheric model gives rise to the cosine-shaped pulse for daytime.

The half-cosine model in the actual GPS ionospheric group delay model is represented by the first three terms in the series expansion $\cos x \cong 1 - (x^2/2) + (x^4/24)$, $|x| < 1.57 \approx \pi/2$, as shown in the algorithm of Table 14.

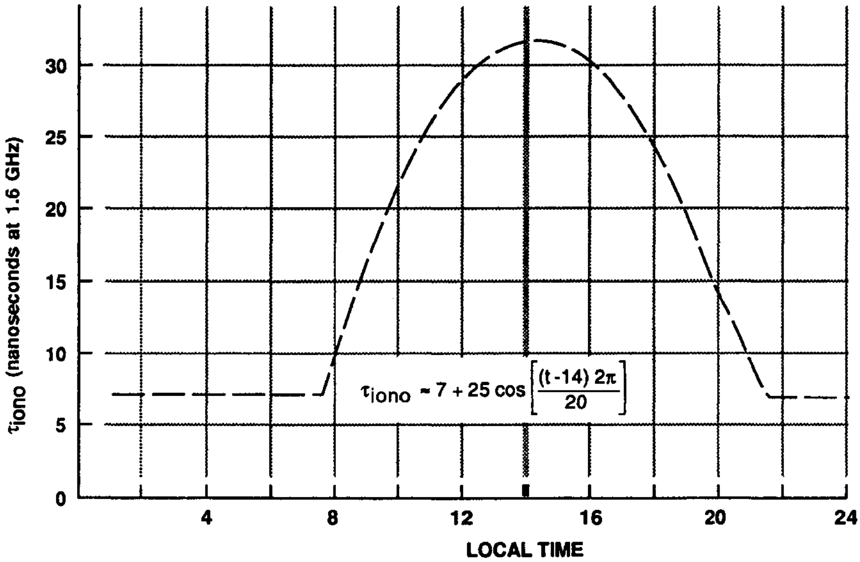


Fig. 8 Example ionospheric group delay cosine fit model.

The specific ionospheric parameters for GPS are given by subframe 4, page 18 for the single-frequency L_1 (or L_2 user) for use in the algorithm given below. These data occupy bits 9–24 of word three plus the 24 MSB of words four and five. The scale factors are shown in Table 15.

Notice that with this definition of the obliquity factor F at elevation angle E = 0-deg gives $F = 3.382032$. Thus, the obliquity factor is slightly greater than three at low elevation angles, and $F = 1.0004$ for $E = 0.50$ semicircles (90 deg).

The GPS ionospheric obliquity factor from the model is plotted in Fig. 9. The obliquity factor of 3.382 at 0-deg elevation angle does not become as great at low elevation angles as the tropospheric obliquity factor, because the ionosphere occupies an altitude range of approximately 50–500 km, and as a consequence of this altitude, even a ray path at 0 deg elevation angle on Earth enters the ionosphere at a steeper elevation angle than 0 deg, whereas the troposphere has maximum effect right at the Earth's surface (see Chapter 13, this volume).

The approximate behavior of the variation of ionospheric delay vs. elevation angle can be estimated by referring to Fig. 10, which shows a uniform spherical shell of ionosphere extending from one altitude h_{min} to an upper altitude h_{max}. The upper and lower limits shown are only examples, the ionosphere is in reality not uniform, and in any event, the upper and lower extent of any model would vary with time. In contrast to the troposphere, the ionosphere does not extend to the Earth's surface. Thus, even if the user elevation angle at the Earth's surface is $E = 0$ deg, the angle of incidence ϕ to the ionosphere for this simple model is greater than zero and for this example: $\phi = \cos^{-1}\{[1/(1 + \delta)]\cos E\} = 13.455$ deg for $E = 0$ deg, where $\delta = (h_{\text{min}}/R)$, and for this example $h_{\text{min}} = 180$ km and $\delta = 0.02822$. The length L_I through the ionosphere for this simple model is easily determined to be as follows:

Table 14 Ionospheric model

The ionospheric correction model is given by the following:

$$I = T_{\text{iono}} = \begin{cases} F^* \, [(5.0 * 10^{-9}) + (\text{AMP})\,(1 - (x^2/2) + (x^4/4))], & |x| < 1.57 \text{ day} \\ F^* \, (5.0 * 10^{-9}), & |x| \geq 1.57 \text{ night} \end{cases}$$

where $I = T_{\text{iono}}$ is referred to the L_1 frequency; if the user is operating on the L_2 frequency, the correction term must be multiplied by $\gamma = 1.646944444$. (T_{iono} is the notation of ICD-GPS-200.)[1]

$$\text{AMP} = \left\{ \begin{matrix} \sum_{n=0}^{3} \alpha_n \, \phi_m^n, & \text{AMP} \geq 0 \\ \text{if AMP} < 0, \text{ set AMP} = 0 \end{matrix} \right\} \text{(s)} \qquad \text{Amplitude}$$

$$x = \frac{2\pi\,(\text{t} - 50400)}{\text{PER}}, \text{ (rad)}$$

$$\text{PER} = \left\{ \begin{matrix} \sum_{n=0}^{3} \beta_n \, \phi_m^n, & \text{PER} \geq 72000 \\ \text{if} & \text{PER} < 72{,}000, \ \text{set PER} = 72{,}000 \end{matrix} \right\} \text{(s)} \qquad \text{Period}$$

$F = 1.0 + 16.0\,[0.53 - \text{E}]^3$, the obliquity factor

α_n and β_n are the satellite transmitted data words with $n = 0, 1, 2,$ and 3.

Other equations that must be solved are as follows:

$$\phi_m = \phi_i + 0.064 \cos(\lambda_i - 1.617) \text{ (semicircles)},$$

$$\lambda_i = \lambda_u + \frac{\Psi \sin A}{\cos \phi_i} \text{ (semicircles)},$$

$$\phi_i = \left\{ \begin{matrix} \phi_u + \Psi \cos A \text{ (semicircles)}, \ |\phi_i| \leq 0.416 \\ \text{if } \phi_i > 0.416, \text{ then } \phi_i = +0.416 \\ \text{if } \phi_i < -0.416, \text{ then } \phi_i = -0.416 \end{matrix} \right\} \text{(semicircles)},$$

$$\Psi = \frac{0.0137}{E + 0.11} - 0.022 \text{ (semicircles)},$$

$t = 4.32 * 10^4 \, \lambda_i + \text{GPS time (s)}$, t = local solar time (s)

where $0 \leq t < 86{,}400$, therefore: if $t \geq 86{,}400$ seconds, subtract 86,400 seconds; if $t < 0$ seconds, add 86,400 seconds.

The terms used in computation of ionospheric delay are as follows:

Satellite Transmitted Terms

α_n = the coefficients of a cubic equation representing the amplitude of the vertical delay, 4 coefficients—8 bits each

β_n = the coefficients of a cubic equation representing the period of the model, four coefficients—8 bits each

(Table 14 continues on next page.)

Table 14 Ionospheric model (continued)

Receiver Generated Terms

E = elevation angle between the user and satellite, semicircles
A = azimuth angle between the user and satellite, measured clockwise positive from the true north, semicircles
ϕ_u = user geodetic latitude, semicircles, WGS-84[5]
λ_u = user geodetic longitude, semicircles, WGS-84[5]

GPS time receiver computed system time

Computed Terms

x = phase, rad
F = obliquity factor, dimensionless
t = local time, s
ϕ_m = geomagnetic latitude of the Earth projection of the ionospheric intersection point, mean ionospheric height assumed 350 km, semicircles
λ_i = geodetic longitude of the Earth projection of the ionospheric intersection point, semicircles
ϕ_i = geodetic latitude of the Earth projection of the ionospheric intersection point, semicircles
Ψ = Earth's central angle between user position and Earth projection of ionospheric intersection point, semicircles

Table 15 Ionospheric parameters from subframe 4, page 18

Parameter	Number of bits	Scale factor, LSB	Effective range[a]	Units
α_0	8^b	2^{-30}	——	s
α_1	8^b	2^{-27}	——	s per semicircle
α_2	8^b	2^{-24}	——	s per semicircle^2
α_3	8^b	2^{-24}	——	s per semicircle^3
β_0	8^b	2^{11}	——	s
β_1	8^b	2^{14}	——	s per semicircle
β_2	8^b	2^{16}	——	s per semicircle^2
β_3	8^b	2^{16}	——	s per semicircle^3

[a] Effective range is the maximum range indicated by the bit allocation and scale factor.
[b] Parameters so indicated shall be two's complement, with the sign bit (+ or −) occupying the MSB.

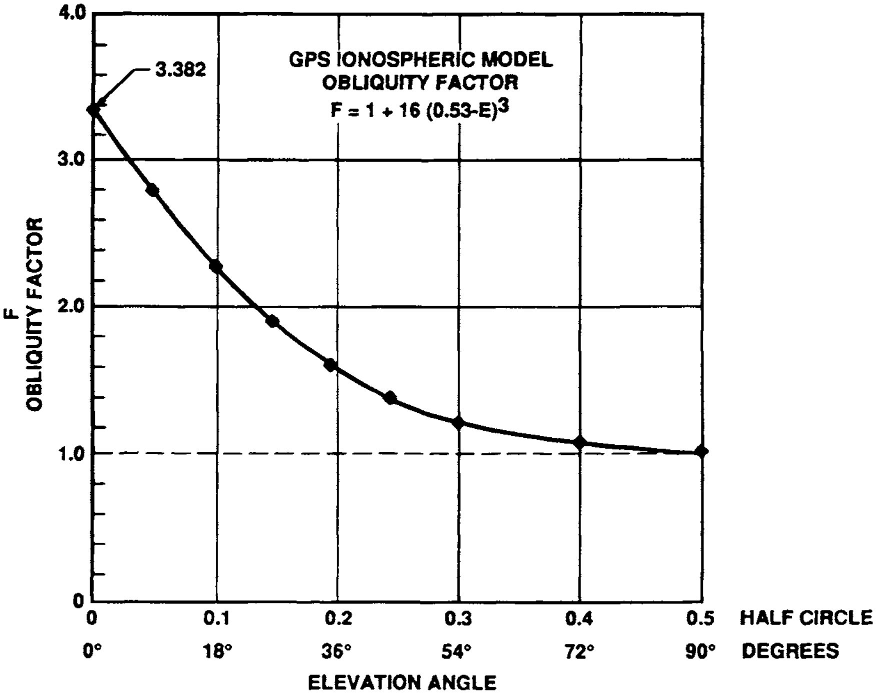


Fig. 9 GPS ionospheric model obliquity factor.

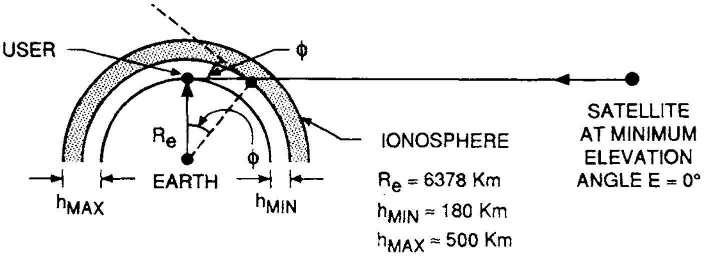


Fig. 10 Approximate representation of the ionosphere by a spherical shell of width $h = h_{max} - h_{min}$. The angle of incidence of the ionosphere *d* relative to the elevation angle E is given by $\phi = \cos^{-1}[(\cos E)/(1 + \delta)]$ where $\delta = h_{min}/R_e$.

Chapter 5

Satellite Constellation and Geometric Dilution of Precision

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

THE GPS satellite constellation is selected to satisfy many different conditions to provide worldwide three-dimensional navigation. The satellite constellation has been placed in an inclined orbit at 55 deg (formerly 63-deg inclination for earlier Block I satellites) to provide full Earth coverage, whereas a purely equatorial orbit cannot provide coverage above approximately 72° latitude at the GPS orbit altitude. The requirement for a minimum of four simultaneous pseudorange measurements with good geometry directly leads to a constellation of many satellites. A nongeosynchronous orbit was selected to permit the use of carrier phase/Doppler measurement profiles in addition to pseudorange measurements of code phase/delay. This objective was one of the reasons for the selection of the 12 sidereal hour orbit.

This chapter discusses the operational satellite constellation of 24 satellites termed the GPS-24 (closely related to the previous Primary 21) constellation. The specific characteristics of this constellation are discussed in terms of the statistics of the numbers of satellites visible at different user latitudes and various elevation angles, the satellite ground tracks, and the signal Doppler shifts. The potential for the future augmentation of this GPS constellation by adding satellites in either the same planes or as geostationary satellites is also discussed.

The chapter then continues with an analysis of the *geometric dilution of precision* (GDOP) and the related DOPs: PDOP, HDOP, VDOP, TDOP for position, horizontal, vertical, and time dilutions of precision. The bounds on the minimum value of these DOPs are analyzed for various numbers of satellites for both two- and three-dimensional problems with various constraints on elevation angle and user clock stability. Further discussion and development of the DOP concept is found in Chapter 11, this volume. The chapter concludes by describing the GDOP for the GPS-24 satellite constellation with a single satellite outage.

*Ph.D., Chairman of the Board.

II. GPS Orbit Configuration, GPS-24

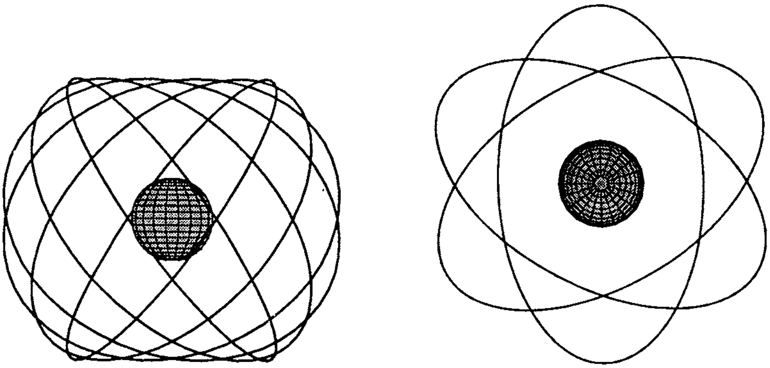
As mentioned earlier, the operational constellation is defined as GPS-24. The selected satellite orbital constellation contains 24 satellites in six orbit planes. There are four satellites in each of the six planes, which are shown in Fig. 1. The satellites have a period of 12 h *sidereal time** and a semimajor axis of 26,561.75 km.† A sidereal day is defined as the time for the Earth to complete one revolution on its axis in Earth–Centered-Inertial (ECI) space and consists of 24 sidereal hours where 1 sidereal day is slightly shorter than a mean solar day (see Chapter 4, this volume). One sidereal day is 23 h, 56 min, 4.009054 s or 86,164.09054 s of mean solar time. One mean sidereal day is equal to 0.997269566 mean solar day. The satellites are inclined with respect to the equatorial plane by 55 deg. Table 1 summarizes the nominal parameters of the GPS orbit.

A. GPS Orbit-Semi-Major Axis

The nominal period T_{po} of an orbit around the Earth is as follows:

$$T_{po} = \frac{2\pi}{\sqrt{\mu}} a_o^{3/2} \quad \text{and} \quad a_o = (T_{po}^2 \mu/(2\pi)^2)^{1/3} \qquad (1)$$

where a_o is the semimajor axis in m, and μ and π are defined for GPS using



a) Viewed from Equatorial Plane

b) Viewed from Pole

Fig. 1 GPS-24 satellite constellation: a) the six orbit planes inclined at 55 deg shown in Earth–centered-inertial (ECI) coordinates viewed from the equatorial plane. The GPS constellation has four satellites in six equally spaced orbit planes. Note that the symmetrical satellite orbit planes are superimposed in part b. The present GPS-24 satellite constellation contains four nonequally spaced satellites in six orbit planes. The satellite phasing is designed to minimize the effect of satellite outage. The satellite phasing is shown in Fig. 2.

*A satellite with a 12-h orbit in sidereal time rotates once in inertial space in 12 h of sidereal time. The semimajor axis for a 12-h sidereal orbit is 26,561.75 km and has an altitude above the Earth equatorial radius of 6378.137 km of 20,183.6 km.

†Strictly speaking, the GPS satellites have a 12-h orbit (sidereal time) wherein the ground tracks repeat to within ± 2° in longitude. If the ground track drifts off from its nominal value by more than 2°, the GPS control segment makes a minor orbit correction.

Table 1 Approximate GPS satellite parameters and physical constants

Parameter	Value	
Orbit plane spacing	6 equally spaced ascending nodes at 120 deg	
Orbit radius r_{cs}	26,561.75 km semimajor axis	
Orbit velocity (circular) (ECI)	$\cong \sqrt{\mu/r_{cs}} = 3.8704$ km/s	
Eccentricity	Nominally zero but generally less than $e = 0.02$	
ω_s angular velocity	$2 \times 7.29211 \times 10^{-5}$ rad/s	
Period	12-h mean sidereal time	
Inclination	$i = 55$ deg nominal	
Velocity of light c	2.99792458×10^{8} m/s	GPS-ICD-200[a]
Earth's gravitational parameter μ	3.986005×10^{14} m^3/s^2	GPS-ICD-200 WGS-84 value[b]
Pi, π	3.141592653898	GPS ICD-200
Earth's rotation rate Ω_e in inertial space	$7.2921151467 \times 10^{-5}$ rad/s	ICD-200 WGS-84 value
One sidereal day $2\pi/\Omega_e$	86164.0989038 s	WGS-84 (calculated from $\dot{\Omega}_e$)
One solar day	86400 s	
J_2 harmonic	1.08268×10^{-3}	

[a] GPS-ICD 200, GPS Interface Control Document.
[b] WGS-84, 1984 World Geodetic System.

WGS-84 and GPS-ICD-200 values as shown in Table 1. The values of a_o and T_{po} neglect the effect of Earth oblateness. If the period of the orbit is set as 1/2 of a sidereal day, (12 h sidereal time), then the nominal value of the semimajor axis a_o is 26,561.765 km. However, because of Earth oblateness, the orbit period must be corrected slightly.*

The orbit mean motion (angular) of the satellite with the J_2 correction for Earth oblateness is as follows[1]:

$$\bar{n} = \sqrt{\frac{\mu}{a^3}} \left[1 + \frac{3}{2} \frac{J_2 R_e^2}{p^2} \left(1 - \frac{3}{2} \sin^2 i \right) (1 - e^2)^{1/2} \right] \tag{2}$$

where a is the corrected semimajor axis at epoch; e = eccentricity; $p = a(1 - e^2)$; R_e is the Earth equatorial radius; i is the orbit inclination; and $J_2 \cong 1082.68 \times 10^{-6}$. For $e \cong 0$, this equation reduces to the following:

$$\bar{n} \cong \sqrt{\frac{\mu}{a^3}} \left[1 + \frac{3}{2} \frac{J_2 R_e^2}{a^2} \left(1 - \frac{3}{2} \sin^2 i \right) \right] \Bigg\} \quad \text{for} \quad e \cong 0$$

$$\cong \sqrt{\frac{\mu}{a^3}} (1 + \Delta) \tag{2a}$$

where $\Delta = (3/2)\, J_2[R_e/a]^2[1 - (3/2) \sin^2 i]$ for $e \cong 0$.

*There is also a correction for solar pressure, which is dependent on the spacecraft size and shape, that is neglected here.

The period of the orbit corresponds to $\bar{n}T_p = 2\pi$, and then is given by the following:

$$\bar{n}T_p = 2\pi, \quad \text{or} \quad T_p = 2\pi/\bar{n} \cong 2\pi\sqrt{\left(\frac{a^3}{\mu}\right)\Big/[1 + \Delta]} \tag{3}$$

for $T_p = 12$ h sidereal time, and $\Delta << 1$. Solve Eq. (3) for the semimajor axis a to obtain the following:

$$a \cong \left[\mu\left(\frac{T_p}{2\pi}\right)^2(1 + \Delta)^2\right]^{1/3} \tag{4}$$

noting that Δ is inversely proportional to a^2.

For the GPS orbit with an orbit inclination of 55 deg and zero eccentricity, $\Delta \cong -6.098 \times 10^{-7}$. Thus, the second-order effect on Δ by the variation of a can be neglected, and the corrected value of the semimajor axis is as follows:

$$\begin{aligned} a &\cong \left[\frac{(1 + \Delta)^2}{(2\pi)^2}\,\mu T_p^2\right]^{1/3} = \left[\frac{\mu T_p^2}{(2\pi)^2}\right]^{1/3}(1 + \Delta)^{2/3} \\ &\cong \left[\frac{\mu T_p^2}{(2\pi)^2}\right]^{1/3}\left(1 + \frac{2}{3}\Delta\right) \quad \text{for} \quad \Delta << 1 \end{aligned} \tag{5}$$

For GPS, the J_2 correction is relatively small and only corresponds to a 0.0108 = km reduction in the semimajor axis.*

Thus, the corrected value of a is 26561.754 km.

B. GPS Orbit–Satellite Phasing

Table 2 lists the orbital parameters for the GPS-24 satellite constellation. The satellites all have a design eccentricity of $e = 0$ and an inclination $i = 55$ deg. The angle Ω is the right ascension of the ascending node measured in inertial coordinates from the vernal equinox. The *lan* is the longitude of the ascending node in ECEF coordinates at the epoch time 7,1,0,0,0 of year 1993. The quantity M is the mean anomaly, which is the satellite phase for a circular orbit. If the 24 satellites were equally spaced in each orbit plane, there would be a 360 deg/4 = 90 deg separation in M.

In Table 2, note that the longitude of the ascending node *lan* is expressed in Earth-centered, Earth-fixed (ECEF) rotating coordinates. The right ascension of ascending node Ω for each of the six planes *A, B, C, D, E,* and *F* as expressed in ECI coordinates is the same for all four satellites in each individual plane and are offset by 60 deg from one another. All of the phase angles and longitudes of the ascending node *lans* (expressed in ECEF coordinates) are different. Thus, there are 24 different ground tracks. Units are in km and deg. The satellite phases in each of the planes are shown in Fig. 2. Notice that in inertial space, the satellite planes are equally spaced in longitude relative to the vernal equinox, but the

*Note that at the earlier Block I satellite inclination of 63 deg, the semimajor axis reduction is 29.29 times as large.

Table 2 GPS-24 satellite constellation

	ID	a	i	Ω	M	ΔM	lan
1	A3	26561.75	55.0	272.847	11.676	103.55	179.63
2	A4	26561.75	55.0	272.847	41.806	31.13	14.69
3	A2	26561.75	55.0	272.847	161.786	119.98	74.68
4	A1	26561.75	55.0	272.847	268.126	106.34	127.85
5	B1	26561.75	55.0	332.847	80.956	130.98	94.27
6	B2	26561.75	55.0	332.847	173.336	92.38	140.46
7	B4	26561.75	55.0	332.847	204.376	31.04	155.98
8	B3	26561.75	55.0	332.847	309.976	105.6	28.78
9	C1	26561.75	55.0	32.847	111.876	100.08	169.73
10	C4	26561.75	55.0	32.847	241.556	129.68	54.57
11	C3	26561.75	55.0	32.847	339.666	98.11	103.62
12	C2	26561.75	55.0	32.847	11.796	32.13	119.69
13	D1	26561.75	55.0	92.847	135.226	100.07	61.40
14	D4	26561.75	55.0	92.847	167.356	32.13	77.47
15	D2	26561.75	55.0	92.847	265.446	98.09	126.51
16	D3	26561.75	55.0	92.847	35.156	129.71	11.37
17	E1	26561.75	55.0	152.847	197.046	130.98	152.31
18	E2	26561.75	55.0	152.847	302.596	105.55	25.09
19	E4	26561.75	55.0	152.847	333.686	31.09	40.63
20	E3	26561.75	55.0	152.847	66.066	92.38	86.82
21	F1	26561.75	55.0	212.847	238.886	103.54	53.23
22	F2	26561.75	55.0	212.847	345.226	106.34	106.40
23	F3	26561.75	55.0	212.847	105.206	119.98	166.39
24	F4	26561.75	55.0	212.847	135.346	30.00	1.46

Units are in km for a, and degrees elsewhere, identification ID = Plane/slot; a = Semimajor axis, size in km; i = Inclination, deg; Ω = Right ascension of ascending node, deg; M = Mean anomaly, deg; lan = Longitude of ascending node, deg; ΔM = Phase difference in mean anomaly to adjacent satellite in the same plane.

satellites themselves in each plane are not equally spaced. In fact, two of the satellites in each plane are spaced by between 30.0 and 32.1 deg. If the remaining two satellites were equally spaced in the remaining 330 deg, that would put the spacing of the other satellites at approximately 330 deg/3 = 110 deg. The actual separation varies from 92.38 to 130.98 deg. The spacing has been optimized to minimize the effects of a single satellite failure on system degradation. As discussed later in the chapter, even with a single satellite failure, the PDOP does not exceed six for more than 25 min/day for a user constraint on satellite elevation angles of 5 deg or more.

II. GPS Satellite Visibility and Doppler Shift

Some important characteristics of the GPS satellite constellation as they relate to user receiver navigation performance are the number of satellites in view and the range of Doppler shifts. It is critical that at least four satellites be in view, but it is highly desirable that five or more be in view at all times. When one

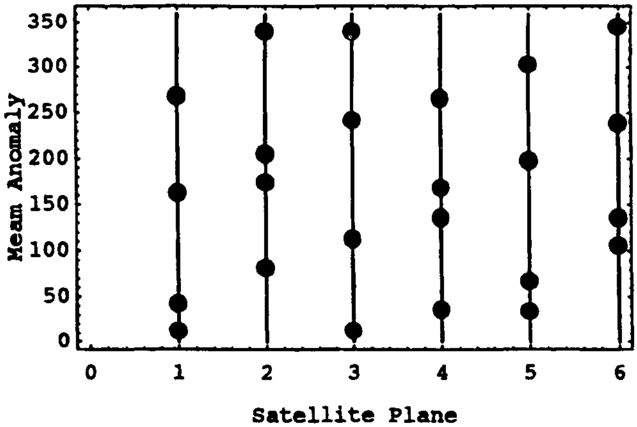


Fig. 2 GPS-24 satellite mean anomaly (satellite phase) in each of the six planes at the epoch 7,1,0,0,0 in 1993 (ECI space). Time is in universal clock time (UTC).

satellite is going out of view, the user receiver must begin to transition to another satellite as its replacement. Furthermore, four satellites by themselves may not provide a sufficiently low GDOP because of poor geometry at certain times.

A. Bound on Level of Coverage for 24 Satellites

A general class of circular orbit satellite constellations with equally spaced satellites and orbit planes has been defined by Walker.[2] In this family of constellations, there are *T* total satellites in *P* uniformly spaced planes of circular satellites, each plane at inclination angle *i* with respect to the equatorial plane. There are *T*/*P* uniformly spaced satellites in each plane. The relative phasing between satellites in adjacent planes is given by *F*, which is in units of 360 deg/*T*; i.e., if a satellite in one plane is just crossing through the equatorial plane in the northerly direction, the adjacent plane satellite is offset by an angle (360 deg/*T*)*F* below the equatorial plane. Thus, the constellation can be described by the notation (*T*/*P*/*F*), and the inclination angle by *i*. For example, if there are $T = 18$ satellites, and the relative phasing between satellites is $F = 2$, then the relative phasing between satellites in adjacent planes as they pass through the equator is (360 deg/*T*)*F* = (360/18)2 = 40 deg.

The degree of worldwide multiple satellite coverage can be determined by computing the maximum Earth angle separation between satellites for a cone that contains *N* satellites. That is, in any conical section of the sphere of satellites there must be *N* satellites in view for a conical angular separation of 2β deg where β deg is determined by the minimum allowed user elevation angle *E* (see Fig. 3). This *N*-satellite visibility must be maintained at all user coordinates and at all times.

Walker[2] has shown that continuous worldwide coverage with at least six satellites in view everywhere is possible with 24 satellites in six planes using a 24/6/1 constellation at an inclination angle of 57 deg for users with a minimum elevation angle of 7 deg. The maximum Earth angle separation between satellites of $\beta = 69.9$ deg for this constellation corresponds to $E \approx 7$ deg minimum elevation

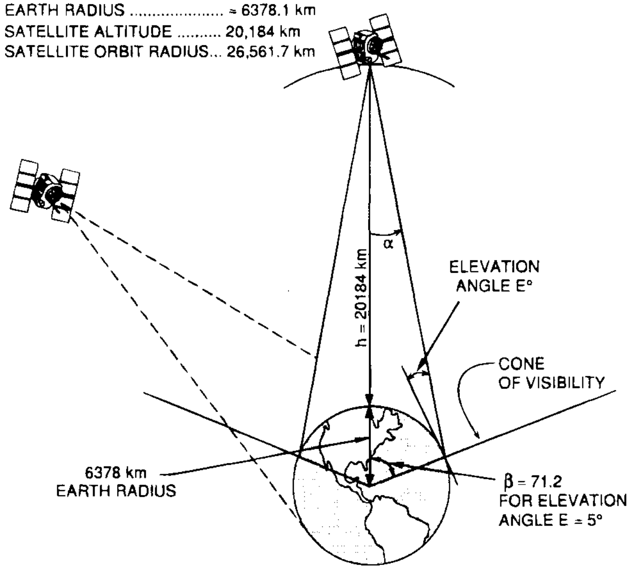

EARTH RADIUS .... ≈ 6378.1 km
SATELLITE ALTITUDE .... 20,184 km
SATELLITE ORBIT RADIUS... 26,561.7 km
α
h = 20184 km
ELEVATION ANGLE E°
CONE OF VISIBILITY
6378 km EARTH RADIUS
β = 71.2 FOR ELEVATION ANGLE E ≈ 5°

Fig. 3 Ground coverage by GPS satellites. The user is restricted to receiving satellites above a minimum elevation angle *E*.

angle (see Fig. 4a). Sevenfold coverage is achievable with a (24/8/4) constellation with a smaller minimum elevation angle of 3 deg is too low for normal operation. However, this constellation's seven-satellite minimum coverage is obtained with the maximum Earth angle separation between satellites of $\beta = 75.8$ deg and the corresponding minimum elevation angle of 3 deg which is too low for normal operation. It should also be noted that the maximum satellite visibility does not necessarily correspond to the smallest GDOP nor to the best performance.

The selected GPS-24 satellite constellation is shown later to give fivefold visibility. Although it does not have as good a full constellation satellite visibility as the (24/6/1) constellation, the GPS-24 satellite constellation has instead been selected on the basis of best coverage if a single satellite becomes inoperative.

B. GPS Satellite Visibility Angle and Doppler Shift

The half angle subtended by the users as viewed from the satellite is termed α as shown earlier in Fig. 3. Each GPS satellite broadcasts to the Earth with an antenna coverage pattern that somewhat exceeds the angle $\alpha = 13.87$ deg subtended by the Earth, as shown in Fig. 3. That is, the satellite antenna pattern extends beyond the edge of the Earth, in order to provide coverage to user satellites not shadowed by the Earth. Each satellite covers users within $\pm\beta$ deg of the subsatellite point, where β is determined by the minimum elevation angle. The subsatellite point can be defined by: A line drawn between the Earth's center, and the satellite intersects the Earth's surface at the subsatellite point. (For a spherical Earth, the satellite is at the zenith of a user at the subsatellite point.) Thus, the fraction of the Earth covered

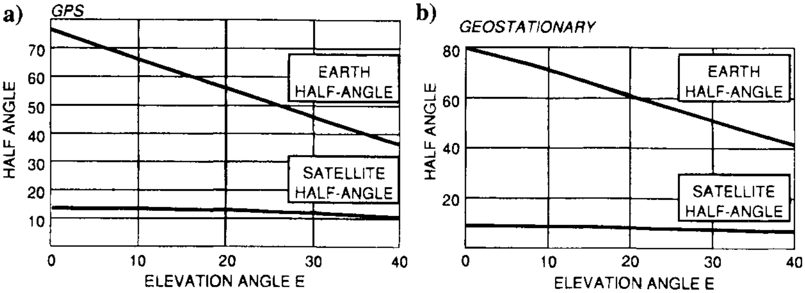


Fig. 4 Half-angles subtended from Earth center β and from the satellite α as a function of the elevation angle *E*. The half-angles are given both for the GPS 12-h orbit (a), and for the 24-h geostationary orbit (b).

depends on the constraint on the minimum elevation angle from the user to the satellite. Within this elevation angle constraint *E*, the value of α is as follows:

$$\alpha = \sin^{-1}\left[\left(\frac{R_e}{R_e + h_s}\right)\cos E\right], \qquad \begin{cases} \alpha = 13.87 \text{ for } E = 0 \text{ deg} \\ \alpha = 13.82 \text{ for } E = 5 \text{ deg} \end{cases} \tag{6}$$

where R_e is the Earth radius; h_s is the satellite altitude; $h_s + R_e = a$; and a is the orbit radius or semimajor axis because we have assumed a circular orbit and spherical Earth.

The Earth half-angle $\beta = 90 \text{ deg} - \alpha - E = 71.2$ deg for $E = 5$ deg, and β is termed the visibility half-angle. Figure 4a shows the visibility half-angles β as a function of elevation angle *E* for GPS altitude orbits. For completeness, the visibility half-angles are shown in Fig. 4b for geostationary 24-h orbits, because there is a consideration of augmenting the GPS orbit satellites with one or more geostationary orbit satellites. The fraction of the spherical shell of satellites visible to the user is $(1/2)(1 - \cos \beta)$. On the average, if all 24 satellites were always equally spaced on a sphere (which they cannot be), we would expect to see a fraction of the satellites corresponding to the fraction of the sphere subtended by the β deg cone; namely, $24(1/2)(1 - \cos \beta) = 8.1$ satellites for $E = 5$ deg of the total 24 satellites.

An example of the actual GPS satellite visibility region for a user at a fixed point on Earth is shown* on a Mercator projection map in Fig. 5. The visibility region is defined by the subsatellite points above the dashed line on the figure.

C. GPS-24 Satellite Visibility

A set of satellite visibility statistics for the operational GPS-24 satellite constellation is shown in Figs. 6 and 7 for various user latitudes. The results have been

*The parametric equation for the locus of latitude and longitude at the edge of satellite visibility (subsatellite point) for a user at latitude ϕ and longitude θ is as follows:

$$\text{latitude} = \sin^{-1}[\sin(71 \text{ deg}) \sin b \cos \phi + \cos(71 \text{ deg}) \sin \phi];$$
$$\text{longitude} = \tan^{-1}[\sin(71 \text{ deg}) \cos \phi / (\cos(71 \text{ deg}) \cos \phi - \sin(71 \text{ deg}) \sin b \sin \phi)] + \theta,$$

where b is the parametric angle $b = (-\pi, \pi)$ and 71 deg is the assumed coverage half-angle from the Earth's center.

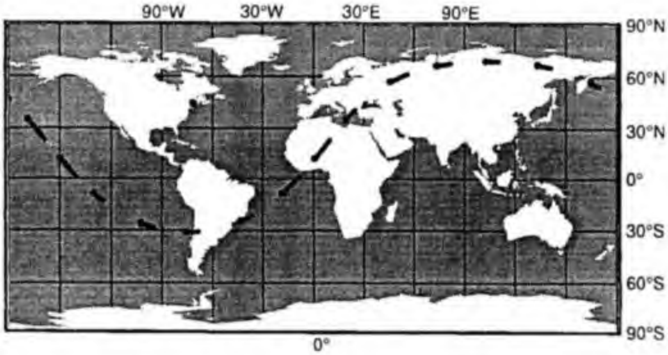
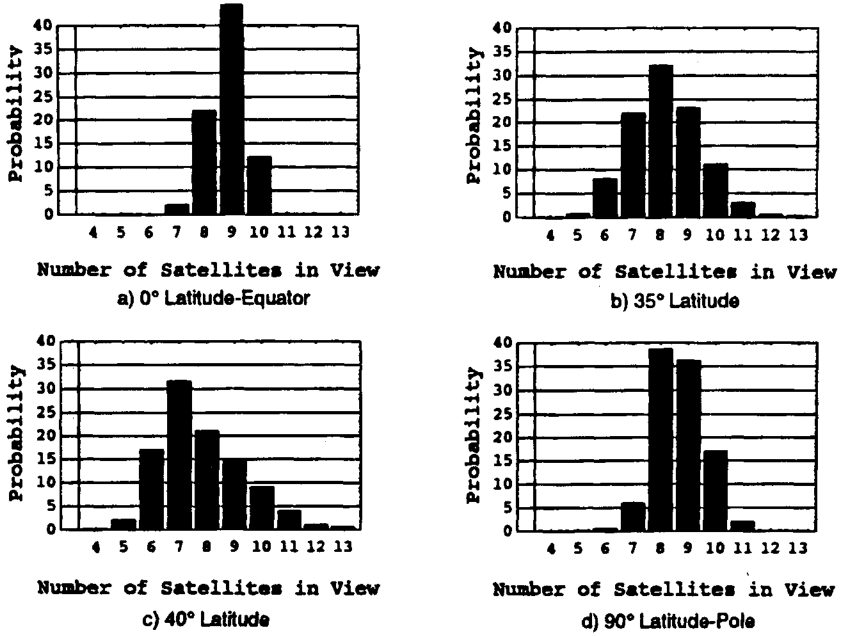


Fig. 5 Example GPS satellite visibility region for a user in the northeastern United States at latitude 40.1°N, longitude 74.5°W (see the dot) with an Earth visibility half-angle of 71 deg. Satellites with a subsatellite point above the dashed curve are in view of the user.



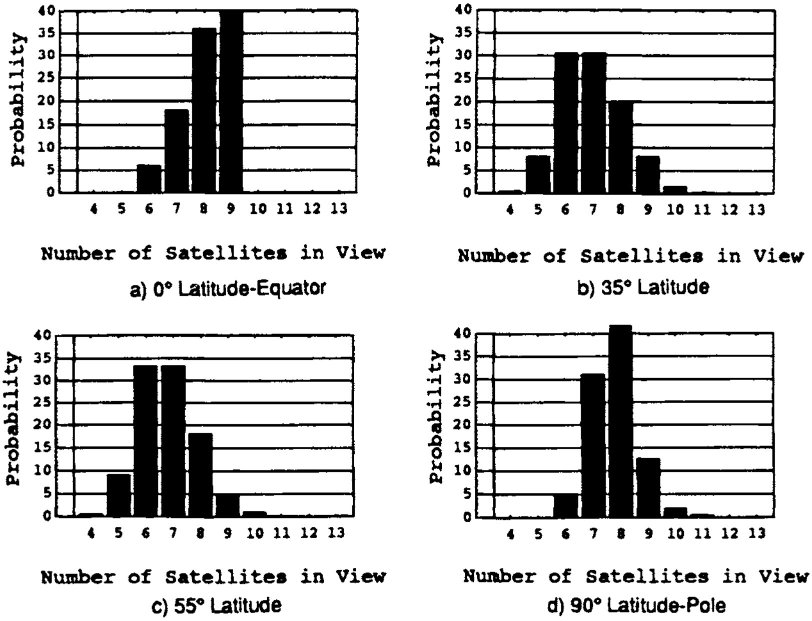
a) 0° Latitude-Equator

b) 35° Latitude

c) 40° Latitude

d) 90° Latitude-Pole

Fig. 6 GPS satellite visibility for the GPS-24 satellite constellation for a 5-deg elevation mask angle. The bar charts correspond to user latitudes as indicated. Note that there are always at least five satellites in view and at least seven more than 80% of the time.



a) 0° Latitude-Equator

b) 35° Latitude

c) 55° Latitude

d) 90° Latitude-Pole

Fig. 7 GPS-24 satellite visibility in percent at various latitudes, averaged over longitude and a 24-h time span. The 35° user latitude corresponds to the approximate worst latitude where momentarily there are only four satellites in view (approximately 0.4% of the time) at this high elevation angle.

averaged over all user position longitudes at that latitude. Figure 6 shows the satellite visibility for a 5-deg elevation mask angle. The majority of the time there are at least seven satellites in view. Compare this result with the previously presented result for equally spaced satellites on a sphere where the average number of satellites on a sphere is 8.1. There are a minimum of five satellites visible for this 5-deg elevation mask angle. Note that a moderately high altitude aircraft can view satellites down to 0-deg elevation angle, and hence have even better satellite visibility. Both sets of results are based on 100 time samples over 24 h and 16 longitude samples at each latitude. It should also be pointed out that a uniform spacing of GPS satellites in an orbit such as the (24/6/1) constellation would lead to better visibility statistics, but greater sensitivity to a satellite outage.

In Fig. 7, a 10-deg elevation mask angle has been used, and satellite visibility has again been averaged over 24 h in time and a set of longitudes at 11.25-deg increments. Note that at both very high and very low latitudes, a minimum of six satellites are always in view and as many as 9–11. However, for a 10-deg elevation mask angle in the vicinity of 35 to 55° latitude, only four satellites are in view a small fraction, <0.5%, of the time. Clearly whenever only four satellites are in view, a failure of one of these satellites would cause an outage. Furthermore, even if four satellites are in view, the satellite/user geometry might correspond to a high GDOP. Thus, although these outages are not very likely and only occur at certain short intervals of time, methods of augmentation of the GPS-24 constellation are of interest.

A word of caution is in order for users at high latitudes. Note that the maximum satellite elevation angle for a user at the pole is

$$E = \tan^{-1}(a \cos (55 \text{ deg})/a \sin (55 \text{ deg}) - R_e) = 44.7 \text{ deg}$$

because the satellite inclination angle is 55 deg. Thus, there are no satellites anywhere at or near the zenith, and the GDOP is degraded, although satellite visibility is good. If the minimum elevation angle is 5 deg, then the usable range of satellite elevation angles is only 39.7 deg.

D. Augmentation of the GPS-24 Constellation

A number of methods can be used to increase the minimum number of satellites in view by adding satellites to the constellation:

1) Add another satellite to each plane to create a (30/6/X) constellation with five satellites in each plane instead of four.

2) Add a ring of GPS satellites at the same altitude but in the equatorial plane. Because the GPS visibility outage regions appear most likely at the midlatitude range 30–50°, a ring of three evenly spaced GPS satellites can add another satellite in view for users anywhere in the midlatitudes. However, sparing is made more difficult by adding yet another plane of satellites.

3) Add several geostationary satellites. These satellites also cover the midlatitude region, and because they are of higher altitude than the GPS satellites, fewer are required for the same degree of coverage. Second, we may be able to put GPS-like payloads on host geostationary communications or weather satellites.

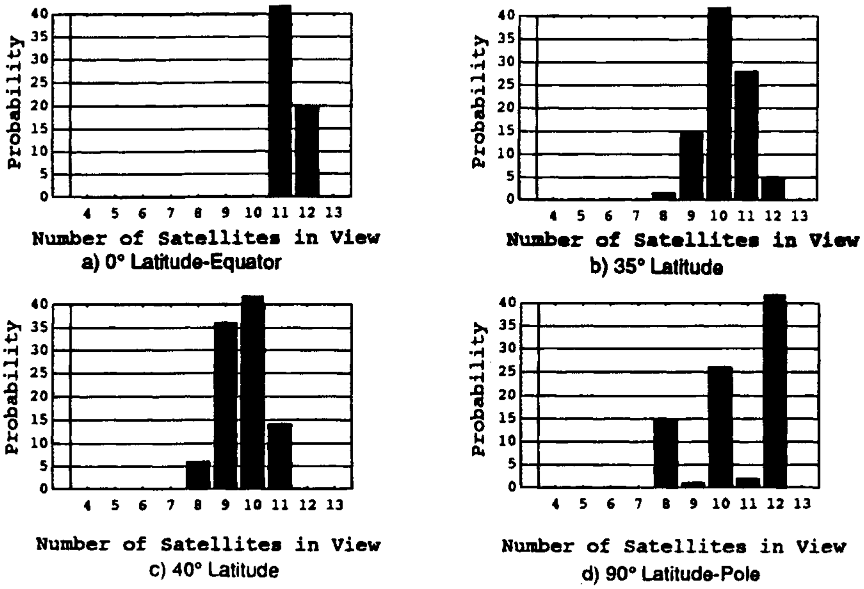
4) Another augmentation being considered is the use of satellites with inclined planes on lower altitude 6-h orbits.

E. Constellation of 30 GPS Satellites

As an example of 30-satellite visibility, the coverage statistics for the (30/6/1) constellation are computed (not necessarily the optimum 30-satellite constellation) and shown in Fig. 8, which illustrates the visibility statistics for a 5-deg elevation mask angle. *Note that here there are a minimum of eight satellites in view at all tested user latitudes.* (55° latitude also showed a minimum of eight satellites in view.) Thus, there seem to be several real advantages in a 30-satellite constellation; namely, visibility is markedly improved; single-satellite outages are easily tolerated; and as is shown later in the chapter on signal-tracking theory, the added power of the other satellites in view can improve receiver performance.

III. Coverage Swath for an Equatorial Plane of Satellites

One of the methods for augmenting GPS is to add an equatorial plane of satellites at either GPS altitude of geostationary altitude. A plane of n equally spaced satellites with separation angle $2\phi = 360/n$ between subsatellite points generates a swath of continuous coverage, as shown in Fig. 9. If the coverage region half-angle for a single satellite for the desired minimum elevation angle



a) 0° Latitude-Equator

b) 35° Latitude

c) 40° Latitude

d) 90° Latitude-Pole

Fig. 8 Visibility statistics for 30 uniformly spaced satellites (30/6/1) constellation vs latitude for a 5-deg elevation mask angle. The user latitudes are as indicated.

E is β as measured from Earth's center, then the coverage swath angle λ in latitude is given by $\lambda = \cos^{-1}[\cos \beta / \cos \phi]$ for satellites in the equatorial plane. If the satellites are at GPS altitude and $\beta = 71$ deg for $E = 5$ deg, then the coverage swath latitudes are as shown in Table 3. Table 3 also gives the "swath for a geosynchronous orbit" at radius 42,162 km.

Thus, it is clear that 3, 4, or 5 satellites in the equatorial plane at GPS altitude produce a continuous coverage swaths of ±49.3, ±62.6, ±66.3 deg, respectively, in latitude about the equator. The operational GPS satellite constellation already has a high degree of redundancy in satellite visibility at the higher latitudes

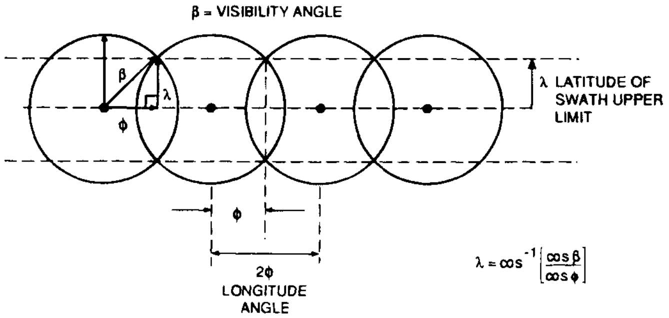

β = VISIBILITY ANGLE
β
λ
ϕ
λ LATITUDE OF SWATH UPPER LIMIT
ϕ
2ϕ
LONGITUDE ANGLE
$\lambda = \cos^{-1}\left[\frac{\cos\beta}{\cos\phi}\right]$

Fig. 9 Visibility coverage swath for *n* satellites in a single plane (equatorial) on a spherical Earth.

Table 3 Coverage swath width latitude λ for $n = 3,4,5$ satellites in an equatorial plane at GPS altitude and geosynchronous altitude (separation $2\phi = 360°/n$)

Number of		GPS		Geosynchronous	
satellites n	Separation 2ϕ	Coverage β	Swath λ	Coverage β	Swath λ
3	120 deg	71 deg	49.4 deg	77 deg	63.3 deg
4	90 deg	71 deg	62.6 deg	77 deg	71.5 deg
5	72 deg	71 deg	66.3 deg	77 deg	73.9 deg
6	60 deg	71 deg	67.9 deg	77 deg	74.9 deg

because the GPS orbits are inclined at 55 deg (see Fig. 6). Momentary periods of low satellite visibility that might occur in the event of a catastrophic satellite failure tend to occur at the midlatitudes, as shown earlier. Thus, it is clear that an equatorial ring of satellites at either GPS or geostationary altitudes with GPS-like payloads could provide an extra degree of redundancy for the GPS system. More specifically, the GPS-24 constellation has momentary periods of only four-satellite visibility for user elevation angle constraints at 10-deg elevation angle and user latitudes in the 30–60-deg region. These regions could be easily covered by three equatorial satellites at either geostationary or GPS altitudes to provide a minimum of five satellite visibility at 10-deg elevation angle or six-satellite visibility at 5-deg elevation angle. Even better would be an additional six satellites in either the (30/6/1) constellation at 55-deg inclination or a set of six equatorial satellites.

A. Satellite Ground Traces

From a user's standpoint, it is sometimes useful to show the ground trace that each of these satellites makes as its subsatellite point moves with time. The ground trace is the line generated on the Earth's surface by the line joining the satellite and the Earth's center as both the satellite moves in its orbit and the Earth rotates. Because the satellites have precisely a 12-h (sidereal time) orbit, each satellite traces out exactly the same track on the Earth's surface each sidereal day.* The Earth, of course, rotates once in inertial space each sidereal day underneath the satellite orbit. Thus the satellite produces a ground track (the locus of points directly below the satellite on the surface of the Earth), which exactly repeats every day. If we stand on a ground track of a satellite, we see the same satellite appear overhead at the zenith every day. In fact, a user at any fixed point sees exactly the same pattern of satellites every day. However, because the user's clock time is mean solar time rather than sidereal time of the satellite period, the user sees this satellite pattern appear approximately four minutes earlier each day (235.9 s).

Define a reference GPS satellite in a circular orbit at inclination i which crosses the equator at $t = 0$ at longitude θ_j. If another satellite of interest is in the same

*As pointed out earlier, the GPS satellite ground tracks are held to within ±2° in longitude by the GPS Control Segment at the time of the writing.

plane but offset in phase (the mean anomaly) by ϕ_i, then it can be shown that this satellite ground track has latitude that varies with time as follows:

$$\text{Lat} = \sin^{-1}[\sin i \sin(\omega t + \phi_i)] \tag{7a}$$

where the period of the orbit $T = 2\pi/\omega$.

The longitude likewise varies with time as follows:

$$\text{Long} = \phi_j - \Omega_e t + \tan^{-1}[\tan(\omega t + \phi_i)\cos I] \tag{7b}$$

where Ω_e is the Earth rotation rate.*

As a simple example, if there were four equally spaced satellites, and one of these crossed the equator at 0° longitude, the ground traces would appear as shown in Fig. 10. Notice that each satellite ground track repeats and makes two complete cycles as it moves from -180 to $+180°$ in longitude, as we would expect from the 12-h period. It can be shown that a satellite in another plane, but offset in phase, can have exactly the same ground track. For example, if one satellite crosses the equator in a positive direction at $t = 0$, another satellite crossing the equator in a negative direction at the same time but 180° offset in longitude would have exactly the same ground track. In general, two satellites have the same ground track if $\Delta\phi = 2\Delta\theta (\text{mod } 2\pi)$. That is, the offset in satellite phase is equal to twice the offset in longitude of their planes modulo 2π. Thus, it is possible for satellites in different orbit planes in inertial space to have the same ground track if the phasing is selected accordingly.

Clearly, with the GPS-24 constellation the satellites do not all have the same ground tracks because the longitude of the ascending nodes (measured in the rotating Earth coordinates) are different (mod π or 180°). Thus, there are 24 separate ground tracks, one for each satellite. However, each of their ground traces have exactly the same shape as those shown here except for the position of the longitude of the ascending node.

IV. Geometric Dilution of Precision Performance Bounds and GPS-24 Performance

The previous section discussed satellite orbits and visibility, and previous and later chapters discuss the pseudorange measurement accuracy. However, it is really user position that is of greatest interest, and as described earlier, the general relationship between the errors in the pseudorange measurements by the user to the user position accuracy is described by the GDOP. The GDOP generally assumes that the measured pseudorange errors are independent with zero mean, and all *measurement errors* have the same rms value σ. (GDOP is also discussed in Chapter 11, this volume, wherein a different approach is taken but leads, of course, to the same result.) The following are the GDOP parameters:

$$\text{GDOP} = \frac{1}{\sigma}\sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2 + \sigma_b^2} \text{ for three dimensions plus time}$$

$$= \text{Geometric Dilution of Precision}$$

*For computational purposes, it is sometimes useful to approximate the $\tan^{-1}(\ )$ relationship for $I = 55$ deg of Eq. (6b) by $\text{Long}(t,\phi,\theta) \cong 2\pi t/12 + \phi - 0.271 \sin 2(2\pi t/12 + \phi) + 0.0367 \sin 4(2\pi t/12 + \phi) - 0.00663 \sin 6(2\pi t/12 + \phi) + 0.00135 \sin 8(2\pi t/12 + \phi) - 2\pi t/2y + \theta)$.

Next Page

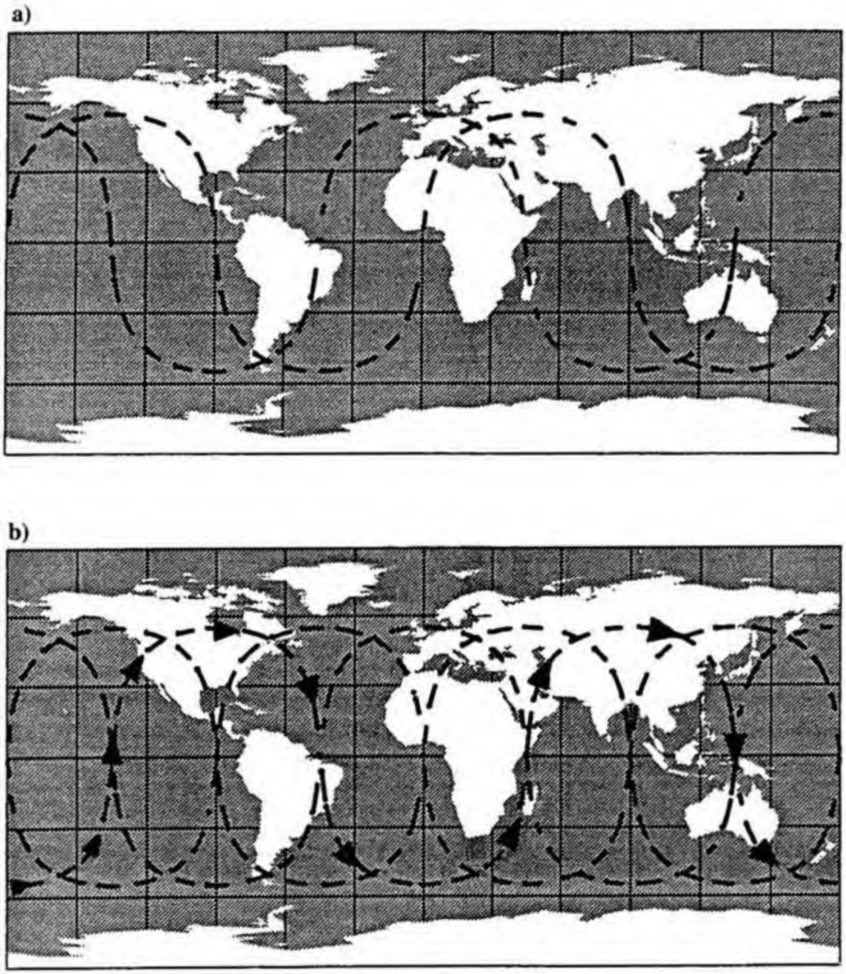


Fig. 10 Example of ground traces for four equally spaced satellites at 55-deg inclination in a single orbit plane at inclination 55 deg where one of the satellites crosses the equator at longitude 0°. Each satellite travels from west to east along its ground track: a) ground traces for two satellites, one of which crosses the equator at 0° longitude (longitude of ascending node is 0°), and the second is in the same orbit plane but spaced by 90 deg; b) ground traces for all four equally spaced satellites. Each satellite follows the ground trace moving from west to east. Note that the GPS satellites are not equally spaced in the GPS-24 constellation.

Chapter 6

GPS Satellite and Payload

M. Aparicio, P. Brodie, L. Doyle, J. Rajan, and P. Torrione*
ITT, Nutley, New Jersey 07110

I. Spacecraft and Navigation Payload Heritage

A. Concept

THE launch of Sputnik I by the USSR in October 1957 was the beginning of the age of navigational satellites. That event was the culmination of centuries of navigation based on the known position of natural heavenly bodies and ever-improving clock accuracy. Observations of the signals transmitted by Sputnik I established the idea that accurate timing signals coming from artificial satellites of known position could aid navigators substantially. In effect, the artificial satellite replaced the functions of the sun and star tables with the known ephemeris of the satellite. In addition, the satellite's accurate timing signals replaced the ship's chronometer. The satellite-based system could provide continuous, worldwide coverage with few satellites; it could be an all-weather system, and it could provide extraordinary position location accuracy.

B. Relation to Earlier Non-GPS Satellites

The challenge of Sputnik I led to the U.S. satellite system Transit I. Begun in December 1958, its goal was continuous, worldwide, all-weather coverage. Transit was placed in operation January 1964. During the late 1960s, the growing need for accurate navigation among the U.S. strategic and tactical forces, combined with the rapid reduction in the cost of computers and processors (user equipment) established the need for, and potential feasibility of, a highly accurate tri-service navigational satellite system. The Air Force and Navy began independent programs called, respectively, 621B and TIMATION. NAVSTAR/GPS is the program that combined features of 621B and TIMATION. The Air Force's signal structure and frequencies were used, while the Navy's orbital configuration was chosen.

C. Overview of Payload Evolution

Two prime contractors under the management of the U.S. Air Force have taken the responsibility of taking the GPS satellite from concept into practice. Rockwell

*Aerospace/Communications Division.

International was the developer of the initial or Block I spacecraft and the Production Block II and IIA spacecraft, and Martin Marietta is the current developer of the GPS Block IIR (replenishment) spacecraft. ITT Corporation developed the critical navigational payload elements on all these satellites. The pseudorandom noise signal assembly (PRNSA), which constitutes the navigation payload on the original eight GPS development spacecraft consists of the following components: baseband processor; L_1/L_2 synthesizer; L_1 modulator; L_2 modulator; L_1 high-power amplifier; L_2 high-power amplifier; and diplexer. It also included delivery of GPS peculiar support equipment (GPSE) for ground station and prelaunch testing. Four more PRNSAs were subsequently added for a total of 12 GPS Block I deliveries.

The development Block I program resulted in 10 GPS spacecraft that were successfully launched from 1978 through 1985. One Block I spacecraft was destroyed because of launch failure. The 12th Block I was a qualification unit that was not flown. The Block II production spacecraft launches began in 1989. The Block II changes to the basic navigation service of the Block I spacecraft consisted of a gradually degrading navigation service for a period of 14 days, if the Control Segment (GPS CS) became inoperable. Block IIA spacecraft added an autonomous momentum management capability that functioned for a period of 180 days without ground contact.

The navigation message uploads to the Block II/IIA spacecraft are performed on a daily basis by the GPS CS. Block IIR satellites incorporate a ranging capability in the satellite crosslink. This ranging capability, combined with an onboard Kalman filter, gives the Block IIR autonomous navigation Autonov capability. Using the crosslink range measurements, the Block IIR spacecraft estimates the error in the Kepler orbital parameters. This enables the Block IIR satellites to support full navigation accuracy of 16-m spherical error probable (SEP), without CS contact for periods of up to 180 days. The comparable error of a Block IIA at the end of 180 days is of the order of kilometers.

D. On-Orbit Performance History

Figure 1 summarizes the historical development of the GPS space segment. Block IIR spacecraft are presently in development and will replace the Block I, II, and IIA satellites as they are declared nonfunctional. There are two basic ways in which satellites become nonfunctional. The first is unplanned and involves an on-orbit failure causing loss of the spacecraft, "caused by," "and" failure of two or more redundant components. The second is depletion of such life-limiting items as thruster fuel and degradaton of the solar arrays or batteries.

The GPS achieved a major milestone in December 1993 when it established initial operational capability or IOC. At IOC, the Air Force Space Command achieved operation of a full constellation of 24 GPS satellites. Shortly thereafter in February 1994, the Federal Aviation Agency (FAA) declared GPS operational for aviation use.

Figure 2 summarizes the launch and on-orbit performance history of the Block's I, II, and IIA space vehicles (SV). The last Block IIA SV is expected to be launched in the first quarter of 1996. Block IIR satellites will begin launches in

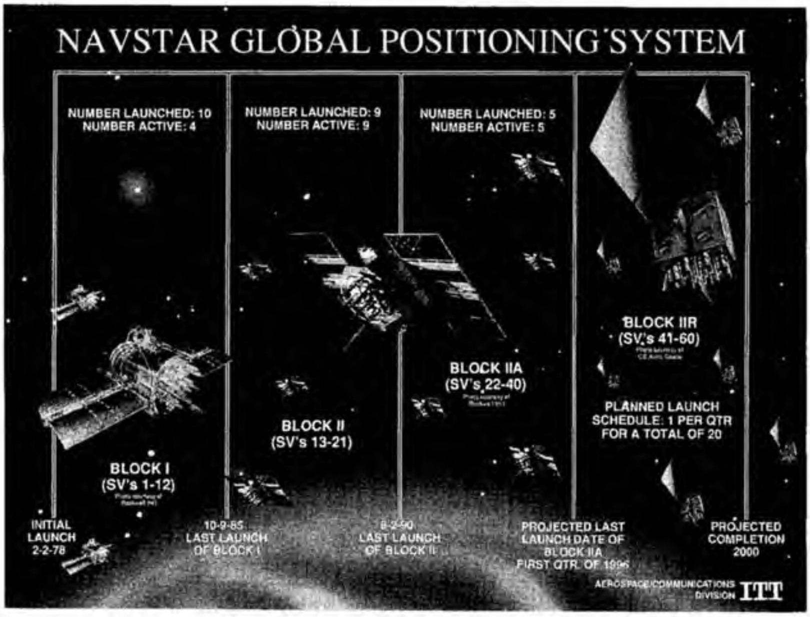


Fig. 1 Historical development of the space segment.

1996. The Air Force plans to begin the Block IIF (Follow-on) program in 1995 with expected launches beginning in 2001.

II. Navigation Payload Requirements

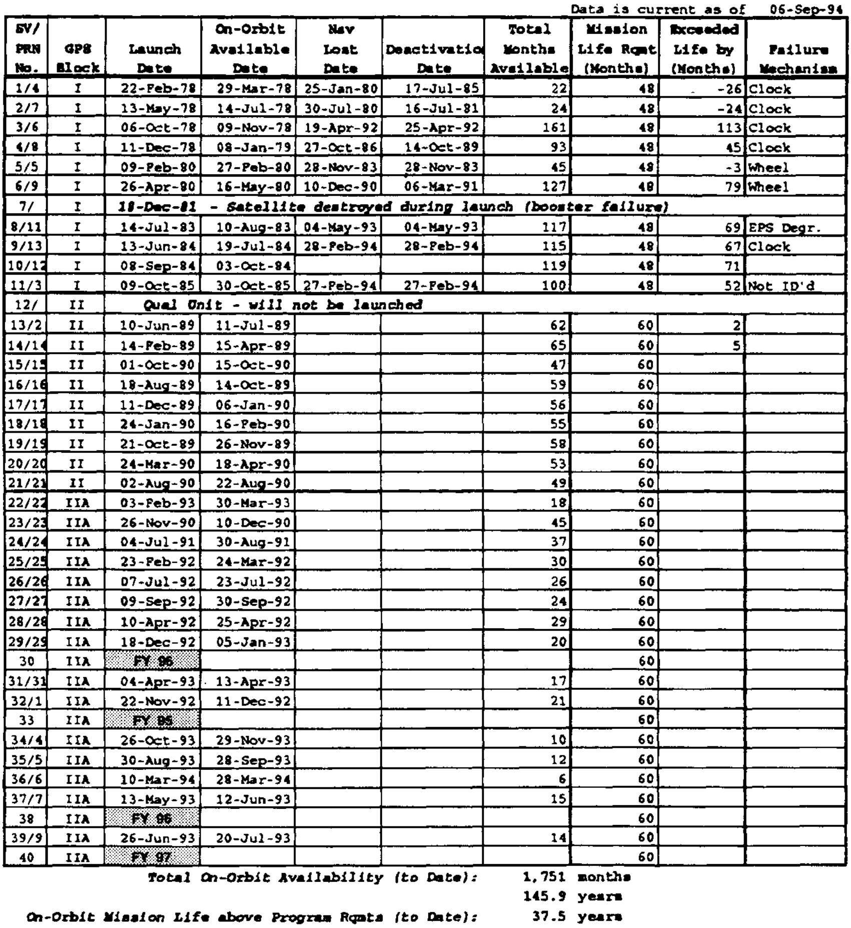
A. GPS System

GPS is a highly accurate, passive, all-weather, 24-h, worldwide, common-grid navigation system. The navigation payload must supply a continuous, precision, high-integrity signal to support this requirement.

The GPS has also enabled specialized users to enhance system accuracy, availability, and integrity with augmentation and sensor integration. The augmentations include differential stations providing differential corrections to GPS measurements; the GPS integrity channel (GIC) providing faster integrity information via geostationary satellites and pseudolites, enabling faster cycle ambiguity resolution and precise localized navigation. The sensor integrations include GPS integrations with inertial navigation systems (INS) and low-cost multisensors to provide accurate navigation comparable to INS at much lower cost. These combined systems provide superior short-term accuracy of GPS and control the error growth during GPS signal outage (caused by masking and aircraft maneuver) by utilizing the long-term stability of an INS.

B. GPS Performance

The GPS is designed to provide a precision positioning service with a 16-m SEP. This performance is achievable by a dual-frequency precision code user



Data is current as of 06-Sep-94

SV/ PRN No.	GPS Block	Launch Date	On-Orbit Available Date	Nav Lost Date	Deactivatio Date	Total Months Available	Mission Life Rqmt (Months)	Exceeded Life by (Months)	Failure Mechanism
1/4	I	22-Feb-78	29-Mar-78	25-Jan-80	17-Jul-85	22	48	-26	Clock
2/7	I	13-May-78	14-Jul-78	30-Jul-80	16-Jul-81	24	48	-24	Clock
3/6	I	06-Oct-78	09-Nov-78	19-Apr-92	25-Apr-92	161	48	113	Clock
4/8	I	11-Dec-78	08-Jan-79	27-Oct-86	14-Oct-89	93	48	45	Clock
5/5	I	09-Feb-80	27-Feb-80	28-Nov-83	28-Nov-83	45	48	-3	Wheel
6/9	I	26-Apr-80	16-May-80	10-Dec-90	06-Mar-91	127	48	79	Wheel
7/	I	*18-Dec-81 - Satellite destroyed during launch (booster failure)*							
8/11	I	14-Jul-83	10-Aug-83	04-May-93	04-May-93	117	48	69	EPS Degr.
9/13	I	13-Jun-84	19-Jul-84	28-Feb-94	28-Feb-94	115	48	67	Clock
10/12	I	08-Sep-84	03-Oct-84			119	48	71	
11/3	I	09-Oct-85	30-Oct-85	27-Feb-94	27-Feb-94	100	48	52	Not ID'd
12/	II	*Qual Unit - will not be launched*							
13/2	II	10-Jun-89	11-Jul-89			62	60	2	
14/14	II	14-Feb-89	15-Apr-89			65	60	5	
15/15	II	01-Oct-90	15-Oct-90			47	60		
16/16	II	18-Aug-89	14-Oct-89			59	60		
17/17	II	11-Dec-89	06-Jan-90			56	60		
18/18	II	24-Jan-90	16-Feb-90			55	60		
19/19	II	21-Oct-89	26-Nov-89			58	60		
20/20	II	24-Mar-90	18-Apr-90			53	60		
21/21	II	02-Aug-90	22-Aug-90			49	60		
22/22	IIA	03-Feb-93	30-Mar-93			18	60		
23/23	IIA	26-Nov-90	10-Dec-90			45	60		
24/24	IIA	04-Jul-91	30-Aug-91			37	60		
25/25	IIA	23-Feb-92	24-Mar-92			30	60		
26/26	IIA	07-Jul-92	23-Jul-92			26	60		
27/27	IIA	09-Sep-92	30-Sep-92			24	60		
28/28	IIA	10-Apr-92	25-Apr-92			29	60		
29/29	IIA	18-Dec-92	05-Jan-93			20	60		
30	IIA	FY 96					60		
31/31	IIA	04-Apr-93	13-Apr-93			17	60		
32/1	IIA	22-Nov-92	11-Dec-92			21	60		
33	IIA	FY 95					60		
34/4	IIA	26-Oct-93	29-Nov-93			10	60		
35/5	IIA	30-Aug-93	28-Sep-93			12	60		
36/6	IIA	10-Mar-94	28-Mar-94			6	60		
37/7	IIA	13-May-93	12-Jun-93			15	60		
38	IIA	FY 96					60		
39/9	IIA	26-Jun-93	20-Jul-93			14	60		
40	IIA	FY 97					60		

Total On-Orbit Availability (to Date): 1,751 months
145.9 years
On-Orbit Mission Life above Program Rqmts (to Date): 37.5 years

Fig. 2 The GPS satellite on-orbit performance history.

under dynamic environment. The current policy is to encrypt the precision code so that it is available only to authorized users. The user ranging error (URE) is derived from the Block II requirement that the system meet a navigational accuracy of 16-m SEP. The derived user ranging error (URE) is 6.6 m (one sigma). The URE budget is shown in Table 1. The velocity accuracy achievable by the system in a dynamic environment is 0.1 m/s. The time transfer accuracy that can be achieved by the system is at least 100 ns. The system also provides a standard positioning service (SPS) of 100-m 2d rms for unauthorized users. The C/A code by itself can provide an accuracy of about 25-m SEP. However, the accuracy is degraded to a 100-m level by man-made degradation of the navigation signal and parameters.

The total root-sum-square errors (rss) allowed for the space segment is approximately 3.5 m (one sigma). An rss is taken, because these errors are uncorrelated

Table 1 System error budget (Block II)

Error sources	User range error (1σ), m
Space segment	
Clock and navigation subsystem stability	3.0
L-Band phase uncertainty	1.5
Predictability of SV parameters	1.0
Other	0.5
Control segment	
Ephemeris prediction and model implementation	4.2
Other	0.9
User segment	
Ionospheric delay compensation	2.3
Tropospheric delay compensation	2.0
Receiver noise and resolution	1.5
Multipath	1.2
Other	0.5
Total (root-sum-square) rss URE	6.6

or statistically independent. Achieved performance for the satellite is 2.2 m (one-sigma) using the rubidium atomic frequency standard (AFS), and 2.9 m (one-sigma) for the cesium AFS. Both performance measures are taken 24 h after the last update from the ground. For normal operations, the ground updates the satellite every 24 h.

C. GPS Signal Structure

The GPS satellites provide precise ranging signals at two frequencies, L_1 and L_2. The satellite transmits precise ranging information using precision (P) code and transmits coarse range using the coarse/acquisition (C/A) code. The C/A code is a Gold code of register size 10, which has a sequence length of 1023. The clock rate of the C/A code is 1.023 MHz, and the code period is 1 ms. The P code is clocked at 10.23 MHz. Each satellite uses a different member of the C/A Gold code family. The P code is over 37 weeks long but is short-cycled on a weekly basis. Different satellites use a different 1-week segment of the P code. The short length of the C/A code allows user equipment that has a low-cost clock with time uncertainty of the order of seconds to search the entire code phase of the C/A code quickly and acquire and track the C/A signal. Tracking the C/A code enables the receiver to demodulate the navigation data. The navigation message is a 50 b/s datastream arranged in 25 pages, each page containing 5 subframes, with each subframe containing 10 words of 30 bits each. The navigation datawords are encoded with (32, 24) Hamming parity providing single error correction and double error detection capabilities. The navigation data on the C/A code has, in addition to the ephemeris and satellite clock correction and double error detection information, the hand-over information that enables the receiver to acquire and track the P code. Specialized receivers with direct time transfer capability can directly search, acquire, and track the P code.

Dual-frequency transmission of ranging signals by the GPS satellite enables user equipment with dual-frequency capability to measure ranges at the L_1 and

L_2 frequencies. These measurements allow the user equipment to accurately compensate for the propagation delay through the ionosphere.

The signal levels guaranteed to the users near the surface of Earth are shown in Table 6-2 for various signal components. The C/A-code power level is 3 dB higher than that of P code to enable fast initial acquisition of the C/A code. The L_2 power is considerable lower, because it typically is not used for initial acquisition. The receivers in general search, acquire, and track the L_1 C/A signal. Based on range measurements made while tracking L_1 C/A, a limited and somewhat slower search is made on the L_2 signal to enable L_2 acquisition and subsequent tracking.

D. Payload Requirements

The link budget covers the three segments of the system and satisfies the power density levels guaranteed to the users. The obvious trade-off in the link budget is the antenna gain vs payload transmit power. This trade-off is made by evaluating the cost, size, power, and weight.

Key requirements are derived from the receive bands of the receivers onboard the satellites and from such specialized frequency bands as a radio astronomy band. An additional requirement minimizes the phase noise of the space-borne L-band transmitters.

Other key performance parameters include: group delay variation on the L-band transmitter chain, which has an impact on the payload URE; the uncertainty in the differential group delay between L_1 and L_2, which has an impact on the accuracy to which the ionospheric corrections can be made; and gain flatness, which influences the symmetry of the code autocorrelation function at the receiver.

As in most complex systems, high reliability is an important factor in the GPS payload design. High reliability provides an assurance that the system is providing an accurate signal to meet the user's navigational needs.

The design of the spacecraft incorporates subassembly and component redundancy; environmental controls; the use of flight-proven, high-reliability piece parts (e.g., transistors, integrated circuits, relays); and proven manufacturing and test procedures and practices. Particular design and test emphasis is placed on ensuring that a failure within a component will neither degrade the performance of the components within the spacecraft nor propagate throughout the system.

Single point failures have been absolutely minimized. Detailed reliability analyses have been performed on all spacecraft elements to determine the effectiveness of design trade-offs. Failure modes, effects, and criticality analyses (FMECA)

Table 2 L-band rf power

	L_1		L_2
	I	Q	I
	C/A	P(Y)	P(Y)/C/A
Output power at the antenna input, dbw	+14.3	+11.3	+8.1
Power near the Earth's surface	−160.0	−163.0	−166.0

have been performed on all elements of the SV. The analyses are performed concurrently with the design efforts, thus producing designs that reflect the analyses conclusions and recommendations. The GPS spacecraft is required to meet a 7.5-year life and is designed for a 10-year life.

III. Block IIR Space Vehicle Configuration

As noted earlier, the Block IIR spacecraft will replace the Block II and IIA spacecraft as they become nonoperational. The contract to develop the Block IIR system was awarded to Martin Marietta (then General Electric Astro Space) and to ITT Aerospace/Communications Division (then ITT Defense Communications Division) in 1987. The team consists of more than 20 subcontractors nationwide. The launch of the first Block IIR spacecraft is planned for the last quarter of 1996. An exploded view of the Block IIR SV is shown in Fig. 3.

The GPS navigation signal is generated and transmitted by the total navigation payload (TNP). The AFS is the heart of the TNP. It provides the precision timing needed to achieve GPS accuracy. The TNP includes two rubidium atomic frequency standards and a cesium atomic frequency standard.

The L-band system consists of the three transmitter chains for three radio frequencies denoted as L_1, L_2, and L_3. The L_1 and L_2 frequencies are used for the navigation mission of the GPS, and the L_3 frequency is used by the nuclear detonation detection system (NDS or Nudet), also located onboard the GPS SV.

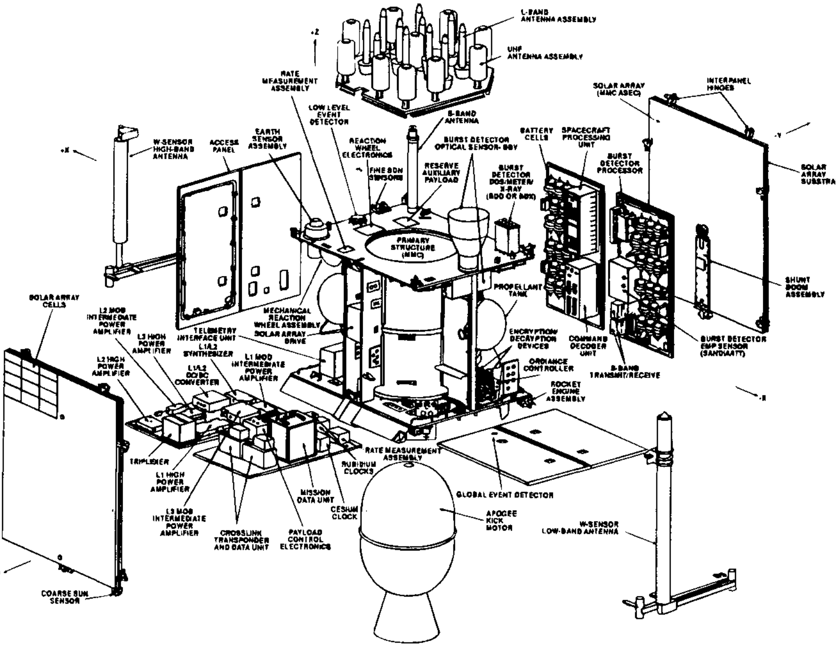


Fig. 3 The GPS Block IIR space vehicle.

Within the TNP, are two major functional capabilities that provide the Block IIR with a giant leap in mission capabilities as compared to Block II/IIA. These are referred to as the time-keeping system (TKS) and the autonomous navigation (AutoNav) capability.

The subsystems of TNP, TKS, and the AutoNav capabilities are explored in detail in the following sections.

The central body of the Block IIR SV is a cube of approximately 6 ft on each side. The span of the solar panel is about 30 ft. The lift-off weight of the spacecraft is 4480 lb and the on-orbit weight is about 2370 lb. The Block IIR spacecraft will be launched by the Delta launch vehicle. The spacecraft has 16 thrusters and a standard set of bus equipment such as telemetry, tracking, and command (TT&C), payload control electronics (PCE), a spacecraft processing unit (SPU), and an attitude reference system. The Block IIR spacecraft are designed to meet 10-year mission life and 4-year storage life. They are designed to be ready to launch with 60-days notice. The SV is designed to operate autonomously for at least 180 days without ground intervention; whereas, the Block I and II satellites' accuracy degrades without ground contact.

A. Navigation Payload Architecture

The TNP components are shown mounted on the two payload panels of the spacecraft shown on the lower left of Figure 6-3. This assembly consists of all the components needed to generate the signals that provide the navigational capability of the GPS. The antennas that radiate the signals to Earth are shown on the top face in the figure. This face is always accurately pointed toward the Earth to ensure uniform and stable illumination of the Earth by the navigational signals. Placing the TNP on two adjacent and stand-alone panels allows easy integration and testing of the spacecraft.

The mission data unit (MDU), atomic frequency assembly, and *L*-band subsystem produce the navigational signals. The crosslink transponder and data unit (CTDU) provides direct satellite-to-satellite communication and ranging. This allows the satellites to operate autonomously (i.e., without ground control segment time and ephemeris updates) at full accuracy for at least 180 days.

The MDU is the brains of the TNP. It integrates all mission functions, such as ephemeris calculations, encryption, NDS data, pseudorandom code generation, and autoNav, as well as monitoring the health of specific TNP components. The MDU software consists of approximately 25,000 lines of code written in Ada running on a MIL-STD 1750A radiation-hard processor at 16 MHz. A block diagram of the TNP is shown in Fig. 4.

IV. Block IIR Payload Design

A. Payload Subsystems

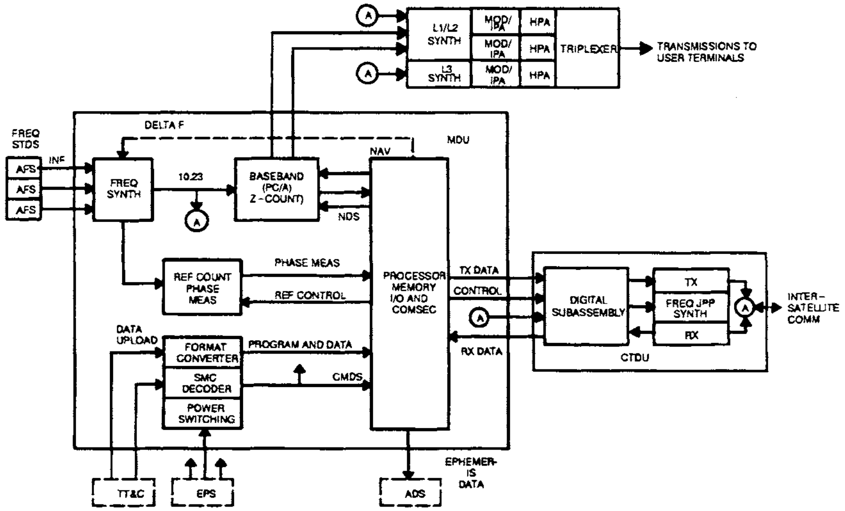
The *L*-band subsystem also includes bi- and quadriphase shift keyed modulators, (BPSK and QPSK) which place the MDU information with the pseudorandom C/A and P codes (1.023 Mbs and 10.23 Mbs, respectively) on the three *L*-band carriers. These three carriers are at 1227.60 MHz (L_2), 1381.05 MHz (L_3), and 1575.42 MHz (L_1). The three carriers are amplified by bipolar transistor

MISSION DATA UNIT

- 1750A Central Processor
- ADA HOL Used Throughout
- Clock Frequency Synthesis from Multiple Standards
- Integral Baseband Processor
- Full Message Encoding and Message Processing
- Real Time Kalman Filter Navigation and Clock State Estimation

CROSSLINK TRANSPONDER DATA LINK

- RF Receive Transmit of Digital Data
- Precision Inter Satellite Ranging
- Frequency Hopped TDMA
- Full Frame Modulation and Mode Control



TIME STANDARD ASSEMBLY

- Multiple Atomic Frequency Standards for Reliability
- Accommodates Various Clock Types (Cs, Rb)
- RAD–Hard Upset Proof Design
- Synthesized High Stability GPS Timing Signals
- Automated Integrity Monitoring

L–BAND SUBSYSTEM

- 25–50 Watt Transmitter
- Bandwidth 20 MHz
- Radiation Hardened
- L_1 : 1 or 10 MChip/s Quadraphase L_2, L_3 : 1 or 10 MChip/s Biphase
- Space Proven Design Operational on Block I and Block II

Fig. 4 Complete system from atomic reference clocks through transmitted navigation messages to users.

amplifiers using microwave integrated circuits technology and sent to the triplexer, which combines them and delivers them to the spacecraft antennas for radiating to users on the Earth. The L_3 signal is filtered through an L_3 astronomy filter.

A principal element of the spacecraft is the atomic frequency standard or atomic clocks. The key requirement is to maintain an accuracy of 6 ns with respect to GPS time. To do this, cesium and rubidium atomic frequency standards are integrated into the TKS of the navigation payload. To ensure operation if one of these critical components fails, two rubidium standards and one cesium standard provide redundancy for the time standard assembly (TSA). The Block

IIR TNP unit is shown in Fig. 5 undergoing prequalification tests at the ITT A/CD facility in Clifton, New Jersey.

The following sections discuss the key design issues associated with each element. Table 3 summarizes the functions of each payload element. Each of these is described in detail in the following sections.

1. Atomic Frequency Standards

a. Introduction. Placing a very stable time reference in a position where maximum user access can be achieved is the basis for modern satellite navigation. At the speed of propagation of electromagnetic signals in the atmosphere, 1 ns of phase uncertainty of the signal, as measured at the user, is roughly equivalent to 1 ft or 1/3 m in position uncertainty. The only instruments that can maintain the phase uncertainty within the required limits for GPS; i.e., 9 ns, one sigma for 24, are atomic frequency standards. This phase stability is equivalent to an uncertainty of 1 mm out of the distance from the Sun to the Earth.

Atomic frequency standards* are amazingly accurate and stable devices. The history of frequency standards and clocks spans the spectrum from sand and water clocks to the modern chronometer and present day AFS. The fundamental problem of clock use has remained the same: "How do we measure the time

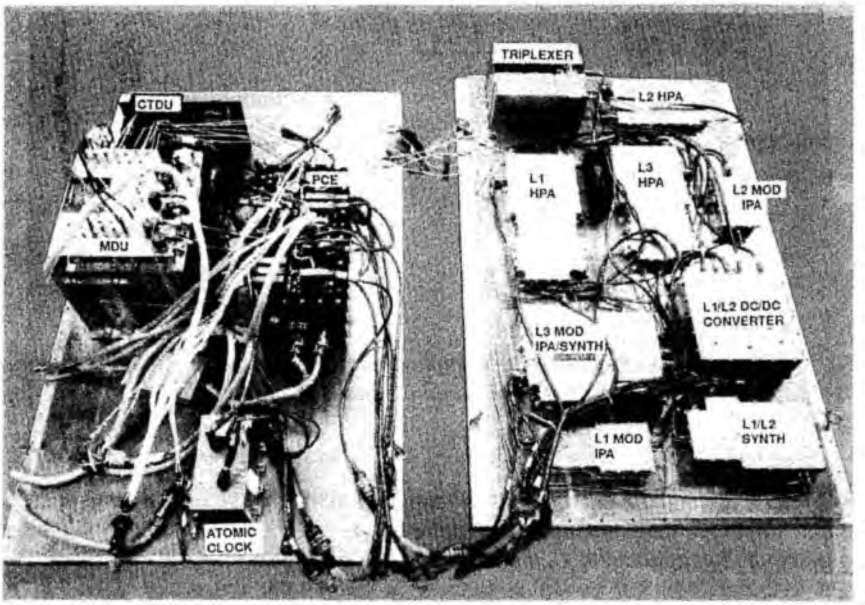


Fig. 5 The GPS Block IIR total navigation payload under test.

*The primary difference between clocks and frequency standards is that clocks keep a record of how many repetitions of a periodic phenomenon have occurred since an event; whereas, frequency standards provide only the periodic phenomenon to the external world.

Table 3 GPS Block IIR payload element functionality

Element	Function
Atomic frequency standards	Provide accurate, stable and reliable timing for all GPS signals. (This is the most important function, because accurate navigation depends primarily upon accurate time.)
Onboard processing	Generate navigational messages, perform ephemeris calculations, and data encryption, generate P and C/A code, monitor health of payload, provide clock error corrections
Software	Implement onboard processing functions. Reprogrammable from ground stations
L-band system	Generate and modulate the L_1, L_2, and L_3 signals and combine them for transmission to Earth
Crosslinks	Provide satellite-to-satellite communications and ranging
Auto navigation	Provide accurate autonomous operation without regular communications from Earth

interval between two events?" The answer has also remained the same; "Count periods of a stable phenomenon." The technological history of clocks and frequency standards is a search for more accurate and stable phenomena and implementation of the measuring device, ranging from the rising and setting of the sun, through mechanical pendulums, to transitions of electrons between energy bands.

The principle of operation of atomic frequency standards[12,14,15] in its most fundamental form is: "The coupling of the output frequency to a periodic natural phenomenon whose period (time to complete a cycle) is essentially invariant." The natural phenomenon used is the change in energy of the outermost electron of an atom of a given element or compound. When this change in energy occurs, the atoms release or absorb the energy at a precisely determined frequency, hence the term atomic frequency standard.

There are different types of AFS, classified by the element or compounds they use for the electron energy transition and the way they couple this information to external devices. Among the most successful AFS types are: hydrogen masers; ammonia frequency standard; cesium frequency standard; rubidium frequency standard and maser; beryllium frequency standard; and mercury electromagnetic ion trap frequency standard.

All AFS types have three common functions: 1) preparation of the outermost electron population into a known state; 2) injection of an electromagnetic signal that causes the energy transition, (higher precision of the frequency of the injected signal causes more atoms to transition); and 3) interrogation and sorting of the resulting energy state of the resulting outermost electron atomic or molecular population. The last step generates an error signal that is used to tune a voltage control oscillator (VCXO) to that natural frequency. The VCXO provides the clock signal with the correct output frequency. Specific examples of this process are the rubidium and cesium AFS. These two types of AFS are of significant importance and are the only types in use in the GPS space segment. The diagrams shown (Fig. 6 and 7) are of the resonator sections of the cesium and rubidium AFS. They are those elements of the AFS that perform the three operations described above.

b. Operation of the cesium AFS. For the GPS cesium AFS, cesium atoms are emitted from a heated cesium reservoir (Fig. 6). Then the three operations are performed. The first is the preparation or sorting of the energy states of the electron in the atomic or molecular population of the cesium atoms. This is accomplished using a magnetic field tuned to the natural frequency of the magnetic dipole of the atoms that contain the electrons in the ground state. This field is produced by a magnet known as the "A" magnet that deflects atoms into the cavity only if they contain electrons in the ground state. This operation creates a relatively pure population of atoms that have their outermost electrons in the ground state.

The second operation is the electromagnetic stimulation of the outermost electron of the cesium atoms. The electrons are shifted from the ground energy state to the next energy state or hyperfine state. This stimulation takes place, if and only if, the electromagnetic field is oscillating extremely close to the specific frequency. That frequency is 9, 192, 631, 770 Hz. If the electromagnetic stimulus is not very close to the oscillating frequency, few transitions to the hyperfine energy state take place, and most electrons remain in the ground energy state. If the electromagnetic stimulus is at the right frequency, many electrons will make the hyperfine transition. This transition changes the properties of the atoms that contain the electrons. In particular, this transition changes the magnetic dipole of the host atoms and allows sorting of the atoms.

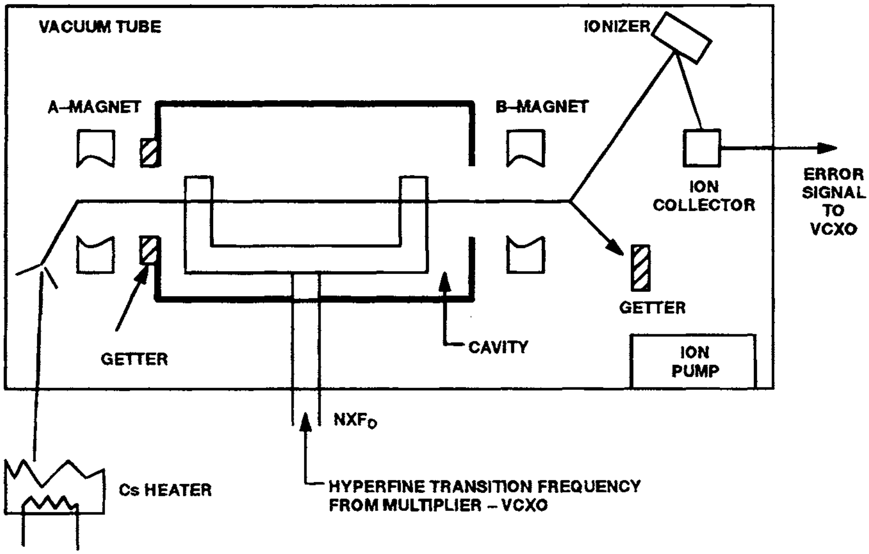


Fig. 6 Cesium beam frequency standard resonator.

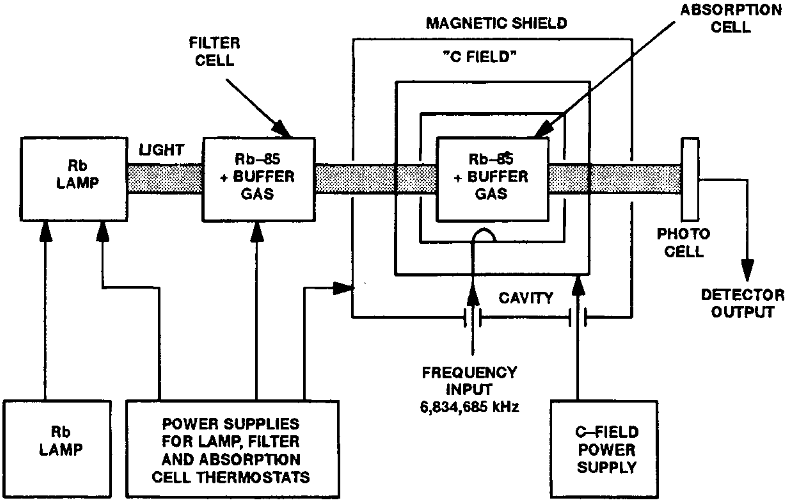


Fig. 7 Rubidium atomic resonator.

The third operation, which is the interrogation and sorting of the atomic population, is performed with the magnet placed at the other end of the cavity. This magnet, the "B" magnet, is tuned so that it deflects atoms that contain electrons in the hyperfine energy state. If the atoms coming out of the cavity contain electrons in the hyperfine energy state, those are then deflected. This means that the electromagnetic stimulus is precisely on frequency if it is successfully causing many energy transitions. The deflected atoms are sent in a path that causes them to strike an ionizer rod. The ionizer rod very efficiently converts the neutral atoms into ions. The ions continue their trajectory into an electron multiplier that turns the ion flow into a current flow proportional to the number of atoms that have been deflected. The current is an indication of the frequency of the electromagnetic stimulus that was injected in the cavity. From the electron multiplier detector output, an error signal is derived, which corrects the VCXO in a control loop. The VCXO oscillation frequency is the actual clock output frequency of the device.

c. Operation of the rubidium AFS. For the rubidium AFS, three similar operations are performed (see Fig. 7). Initially, electrons are in the ground energy state. There will always be a fraction of the total population of atoms that naturally contain outermost electrons in the ground state. In a working standard, this is normally 0.1% of the total population.

The outermost electrons are then excited to the hyperfine energy state. This excitation is carried out by injecting an electromagnetic signal into the cavity that contains the rubidium vapor (the so-called absorption cell). If this stimulus is oscillating at precisely the right frequency 6, 834, 682, 608 Hz (for a zero magnetic field), many electrons in the ground energy state in the atomic population change into the hyperfine energy state. This transition in energy state changes the optical absorption properties of the host atoms.

The third operation is interrogation of the atomic population for energy changes. This optical interrogation process begins at the lamp. The lamp is a discharge device filled with the rubidium isotope 87 and excited by an rf source at about 100 MHz. The lamp emits a spectrum of electromagnetic frequencies, the majority in the visible range. The electromagnetic emissions are passed through a filter cell filled with the rubidium isotope 85. The filter cell stops all but two visible frequencies from reaching the absorption cell. Atoms containing the outermost electrons that have made the hyperfine energy state transition from the ground state will absorb the light. The absorption of the light by the host atoms creates a minimum current output at the optical detector located at the other end of the absorption cell. The minimum occurs only if the electrons in the population of host atoms make the hyperfine energy transition. The optical detector or photodetector generates a current proportional to the amount of light that reaches it. The current is used to shift the VCXO oscillation frequency to the desired clock output frequency.

d. Performance. The most important requirements for operating an AFS in a space-borne platform for a navigational application are: phase stability; reliability; low power; low weight; low volume; and high tolerance to the space environment. To meet these requirements, special space-qualified units are used in the GPS satellites. The most difficult requirements to meet are phase stability and reliability. Although the AFS units used by the National Bureau of Standards exhibit excellent performance, they are too large and heavy for spacecraft applicaton. The reliability of AFS has traditionally been low because of the nature of their high sensitivity to the environment and their complexity. Balancing these unique requirements and constraints to produce units accurate and reliable enough for space poses unique technological challenges.

Phase stability is measured in terms of the time-averaged integral of the fractional frequency stability ($\sigma_y(\tau)$). The fractional frequency stability is measured by the Allan variance $\sigma^2{}_y(\tau)$, which is the square of the Allan deviation $\sigma_y(\tau)$. The Allan variance is defined by the following equation:

$$\sigma_y^2(\tau) = \frac{1}{2N-1} \sum_{i=0}^{N-1} (Y_{i+1} - Y_i)^2$$

where Y_i = fractional frequency at time interval i (actual frequency/nominal desired frequency, and N = number of fractional frequency samples.

The Allan deviation is a measure or statistical estimate of the noise contribution to the frequency instability of frequency standards. The noise processes that have experimentally been found to have significant effects in precision frequency standards are: white phase modulation noise (PM); white frequency modulation noise; flicker phase modulation noise; flicker frequency modulation noise; and random walk of phase noise.

Table 4 is a comparison of some of the critical parameters for the GPS Blocks I, II, IIA, and IIR.[15] As mentioned previously, reliability, power, weight, volume, and tolerance to the space environment are the critical parameters for AFS in the GPS context.

Table 4 GPS Blocks I, II, IIA, and IIR atomic frequency standards specifications

AFS type	Reliability requirement	Size, in.	Weight, lb	Power consumption, W

Rb GPS Block I	0.763 for 5.5 years	L = 5.00
		W = 6.00
		H = 7.50
			13	24.75

Cs GPS Block I[a]	0.663 for 5.5 years	L = 5.30
		W = 15.10
		H = 7.80
			28	22.00

Rb GPS Block II	0.763 for 5.5 years	L = 5.00
		W = 6.00
		H = 7.50
			13	24.75

Cs GPS Block II	0.663 for 5.5 years	L = 5.30
		W = 15.10
		H = 7.80
			28	22.00

Rb GPS Block IIA	0.763 for 5.5 years	L = 5.00
		W = 6.00
		H = 7.50
			13	24.75

CS GPS Block IIA	0.663 for 5.5 years	L = 5.30
		W = 15.10
		H = 7.80
			28	22.00

CS GPS Block IIA second source	0.750 for 5.5 years	L = 5.30
		W = 15.10
		H = 7.80
			28	22.00

Rb GPS Block IIR	0.763 for 7.5 years	L = 8.50
		W = 5.60
		H = 6.20
			14	15

Cs GPS Block IIR	0.775 for 7.5 years	L = 16.50
		W = 5.50
		H = 5.25
			22	26

[a]Only the last four satellites of Block I carry cesium atomic frequency standards.

Table 4 shows that the greatest improvements from Block I, II, and IIA to IIR are the reliability, weight, and power consumption of the rubidium standards. Of these parameters, the most relevant change is the power consumption from 24.75 W to 15 W. Predicted reliability has been increased significantly for both standards. The reliability parameter is of extreme importance in the space mission environment, where the life of components and subsystems limit the mission's life.

B. Mission Data Unit

1. Onboard Processing

Onboard processing is performed in the MDU. The MDU and the frequency synthesizer unit (FSU) are housed together in one physical package. The MDU provides storage of navigation data as uploaded by the CS via the TT&C subsystem of the SV. The MDU combines these data with internally generated ranging codes and sends the resulting navigation message to the *L*-band system (LBS) for transmission to the ground. The MDU also is capable of altering these navigation signals when necessary to deny full GPS navigation accuracy to unauthorized users [i.e., it provides selective availability (SA)]. The FSU plays a role in both the generation of the navigation signals and selective availability, and it is controlled by the MDU processor. The MDU has the capability of autonomously operating for 180 days without receiving navigation data updates from the CS. When the MDU operates in this mode, it computes the SV ephemeris and clock correction data by processing rf ranging performed between spacecraft. The MDU then updates the contents of the navigation message sent to the users.

In addition to the function of providing navigation data to GPS users, the MDU receives data from the NDS. These data are encoded by the MDU and transferred to the ground with navigation data via the LBS. The data are also transmitted via the uhf crosslink to other in-view satellites for retransmission to the ground via the *L*-band. The MDU/FSU operates in the space radiation environment and can operate through or recover from specified transient nuclear events without a permanent impact on navigation functional accuracy or the NDS data-processing function.

The following is a summary of MDU/FSU functions:

1) provide storage of messages uploaded by the CS;

2) process, format, and generate navigation data;

3) provide precise timing for other payload components;

4) generate pseudorandom noise (PRN) codes for navigation;

5) provide SA; i.e., alter the navigation downlink when necessary to deny full GPS navigation accuracy to unauthorized users;

6) perform antispoof (AS); i.e., alter the navigation downlink on CS command to allow the authorized user full GPS accuracy through hostile environments;

7) operate through certain specific transient nuclear events without a permanent impact on the navigation functional accuracy;

8) recover from nuclear radiation transient including logic upset without assistance;

9) autonomously operate for 180 days without update from the CS; i.e., modify navigation data on a periodic basis by processing ranging data from other in-view satellites and exchange navigation data with other SVs and enable autonomous determination of ephemeris and clock corrections;

10) operate for 14 days, with at least the same navigation accuracy as the Block II vehicles are required to meet, without updates from the CS or other satellites;

11) encode NDS data received from the burst detection processor (BDP).

12) insert current SA data and SV ephemeris data into the NDS datamessage for transmission to other GPS satellites and the nuclear detonation user segment (NDUS).

13) update the SA data element (rapid turn-on) via the CTDU upon ground command.

14) perform a graceful turn-on and turn-off function to a known acceptable condition.

15) provide telemetry, diagnostics, and self-check capabilities.

2. *Software*

The computer program in the Block IIR navigation payload is large and complicated, as compared to those used in other spacecraft. This program is referred to as the mission processor (MP) software. The requirement to function autonomously for 180 days dictates that the spacecraft will have to perform many functions currently performed by the CS. In addition to ephemeris and clock parameters estimation, this includes integrity monitoring, curve fitting of the navigation parameters, user range accuracy (URA) estimation, formatting the navigation message, selective availability, universal coordinated time (UTC) steering, and unassisted recovery from upsets.

Another factor contributing to the complexity of the software is the variety of hardware and software interfaces that the processor must deal with concurrently. These interfaces are shown in Fig. 8. Abbreviations for the interfaces that connect directly to the MP are defined in Table 5. Each of these corresponds to one or more I/O ports.

The interfaces connect the MP to the other subsystems shown in Fig. 8, including the SPU; reserve auxiliary payload (RAP); TT&C; CTDU; BDP; FSU; LBS; AFS; hop sequence generator (HSG); COMSEC; watchdog monitor; and error detection and correction (EADC).

The MP is programmed entirely in Ada. This affected the overall approach to developing the software. The Ada tasking model is used to accommodate the wide variety of processing deadlines. Ada portability and modularity are exploited to allow testing of the code in a variety of test environments.

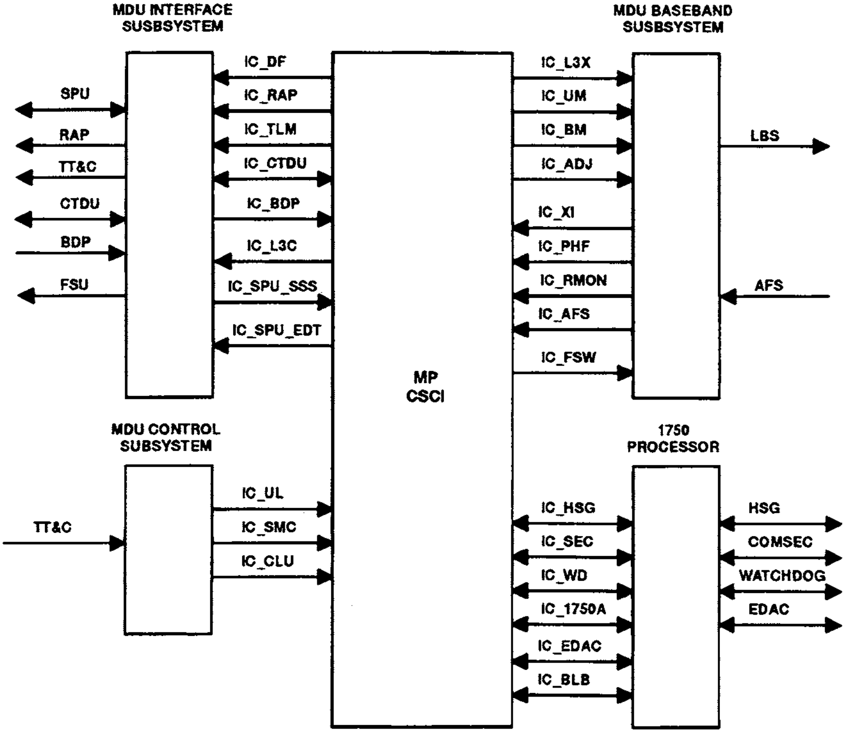


Fig. 8 CSCI–HWCI interfaces.

Table 5 Mission processor CSCI external Interfaces

PUID	Name
IC-UL	Upload data interface
IC-SMC	Serial magnitude command interface
IC-CLU	Clear upload interface
IC-TLM	Telemetry interface
IC-CTDU	CTDU interface
IC-BDP	BDP data interface
IC-L3X	L3 transmit data interface
IC-L3C	L3 on/off command interface
IC-BM	Baseband modulation interface
IC-UM	User message interface
IC-ADJ	Adjustment interface
IC-XI	X1/Z-count interface
IC-PHF	Phase feedback interface
IC-RMON	Reference monitor interface
IC-DF	Delta F command interface
IC-AFS	Atomic frequency standards interface
IC-FSW	AFS switch interface
IC-BLB	Blind bus interface
IC-WD	Watchdog monitor interface
IC-HSG	Hop sequence generator interface
IC-SEC	Comsec interface
IC-1750A	MIL-STD-1750A processor interface
IC-SPU-SSS	Spacecraft subsystem status interface
IC-SPU-EDT	Ephemeris data/time interface
IC-RAP	RAP interface
IC-EDAC	Error detector and correction interface
IC-LLED	Low-level event detector interface

3. *In-Space Reprogrammability*

The operational flight code can be completely reprogrammed from the ground. Upon cold start, the processor executes a program that is resident in PROM (programmable read-only memory). This program then uplinks the operational program over the S-band datalink. The PROM program has sufficient diagnostic capability to verify proper operation of the processor, memory, and data interfaces needed to upload and execute the flight program.

In addition to the capability to upload the entire program, partial uploads are possible. Certain parts of the program have been segmented so that they can be modified without uploading the entire program.

4. *Software Partitioning Facilitates Test*

An important feature of the onboard software is its ability to test certain components in accelerated time. This feature is most significant for AutoNav. A special test bed, the autonomous navigation emulator (ANE), has been developed by ITT to support this testing both before and after launch.

Chapter 7

Fundamentals of Signal Tracking Theory

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

A. GPS User Equipment

THE GPS basic concepts, signal structure, navigation data, satellite constellation, satellite payload, and geometric dilution of precision (GDOP) concepts have all been introduced in previous chapters. This chapter and the next two discuss the user equipment. These chapters provide the foundation for much of the material in the companion volume on GPS applications.

Figure 1 illustrates the basic configuration of a GPS user equipment in its typical form. Generally, the user equipment performs two functions: the first, track the received signals, usually with some form of delay lock loop (DLL) so as to measure the pseudorange and usually the pseudorange-rate or accumulated delta range (ADR), a carrier measurement, as well. This chapter discusses the fundamentals of signal tracking theory as applied to the GPS signals. The next chapter, Chapter 8, discusses means for implementation of the GPS receiver functions. The following chapter, Chapter 9, discusses the navigation algorithm used to convert these receiver measurements into the desired output; namely, user position, velocity, and user clock bias error. This present chapter on signal tracking theory also describes a technique for combining the two functions, receiver tracking and navigation algorithm into one combined integrated tracking system.

In this chapter, the received signals are assumed to be received with stationary additive white Gaussian noise (AWGN) that is representative of the thermal noise at the receiver frontend. For most of this chapter, the received radio frequency signal-to-noise ratio is small, typical of most GPS receiver applications. Thus, the effects of self-noise or multiple access noise caused by the other GPS signals is generally small and is usually neglected. However, in the last section on the vector delay lock loop (VDLL), it is shown that the quasioptimal detector that tracks all signals is designed to remove much of this multiple access noise.

*Ph.D., Chairman of the Board.

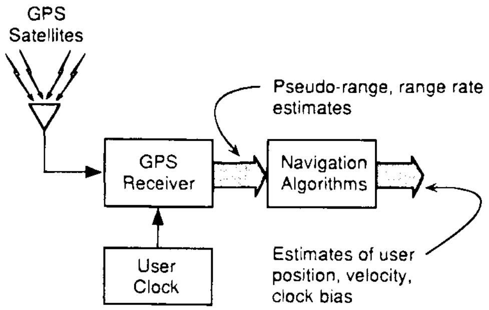

GPS Satellites
Pseudo-range, range rate estimates
GPS Receiver
Navigation Algorithms
User Clock
Estimates of user position, velocity, clock bias

Fig. 1 Simplified configuration of the GPS user equipment.

B. GPS User Equipment-System Architecture

A generalized view of a basic GPS user equipment system is depicted in Fig. 2. The satellite signals are received by one or more antennas/low-noise amplifiers.* The output of the antenna is fed to a radio frequency bandpass filter/low-noise amplifier combination in order to amplify the signal and to filter out potential high-level interfering signals in adjacent frequency bands. Otherwise, these potential interfering signals might either saturate the amplifier or drive it into a nonlinear region of operation. The radio frequency filters must be selected with low loss in order to maintain a low noise temperature and also must have sufficient bandwidth and phase linearity to minimize the distortion of the desired C/A- or P(Y)-code signals. The signal then passes through serial stages of radio frequency amplification, downconversion, IF amplification and filtering, and sampling/quantizing. The sampling and quantizing of the signal can be performed either at intermediate frequency (IF) or at baseband. In either approach, in-phase and quadrature (I, Q) samples are taken of the received signals plus noise. At the present state of the art, we can implement the functions of radio frequency amplification, downconversion, IF amplification, and A/D sampling with a single monolithic microwave integrated circuit chip (MMIC). Filtering and reference frequency generation may require additional circuitry.

The I, Q samples are then fed to a parallel set of DLLs each of which tracks a different satellite, measures pseudorange, and recovers the carrier which is biphase modulated with the GPS navigation data.† The DLL 1–3 and associated

*Although most receivers employ only one antenna, some use more than one antenna/amplifier in order to:

- Accommodate maneuvering of the user platform; e.g., an aircraft banking and thereby avoid blocking some of the satellites with a wing. For example, one antenna can be at the top of the aircraft and others can be at the sides.
- Provide increased antenna gain. Each higher gain antenna can be pointed at single or clustered groups of satellites. The antenna beams can be steered electronically or mechanically, if necessary.
- Discriminate against interfering signals or multipath. Multiple narrow beam antennas or adaptive antennas can be employed. Some antennas use special ground planes to reduce multipath. Null steering antennas can be employed.

†An alternative time sequencing approach can be employed in which a DLL is sequenced over several satellites, dwelling on each satellite for a short period of time. There is, however, some performance degradation with this sequencing approach.

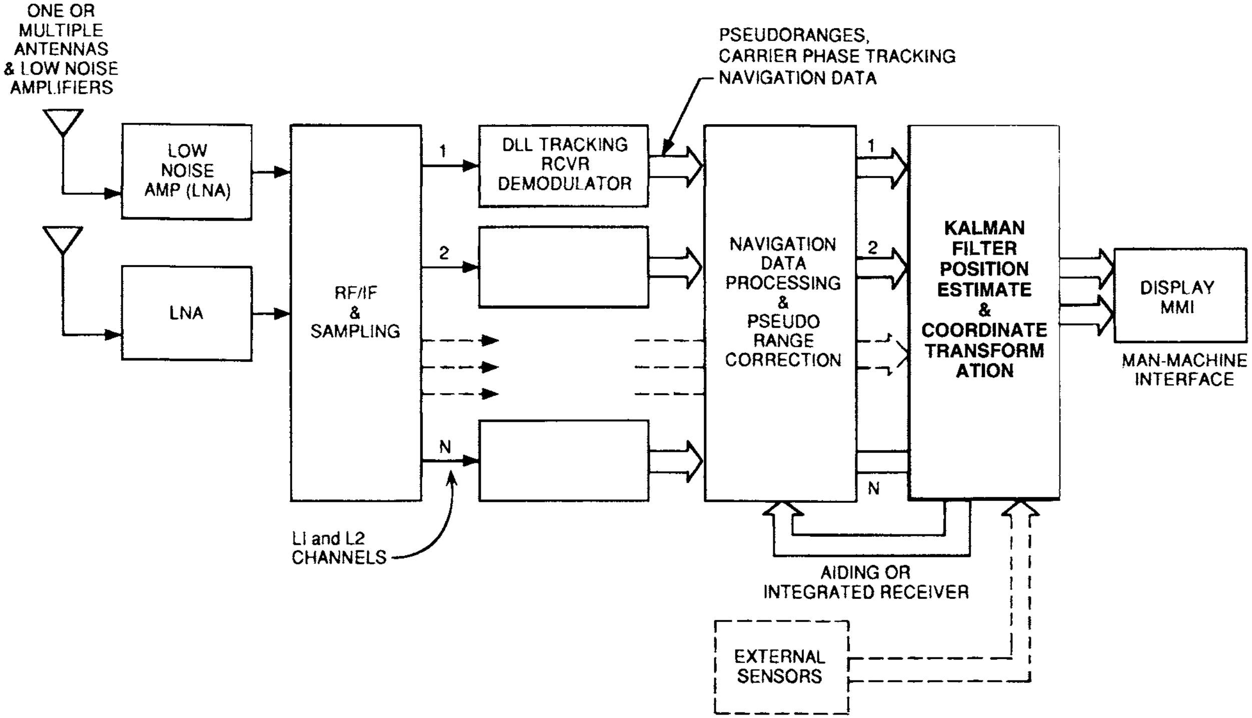

ONE OR MULTIPLE ANTENNAS & LOW NOISE AMPLIFIERS
LOW NOISE AMP (LNA)
LNA
RF/IF & SAMPLING
1
2
N
Ll and L2 CHANNELS
DLL TRACKING RCVR DEMODULATOR
PSEUDORANGES, CARRIER PHASE TRACKING NAVIGATION DATA
NAVIGATION DATA PROCESSING & PSEUDO RANGE CORRECTION
1
2
N
KALMAN FILTER POSITION ESTIMATE & COORDINATE TRANSFORM ATION
DISPLAY MMI
MAN-MACHINE INTERFACE
AIDING OR INTEGRATED RECEIVER
EXTERNAL SENSORS

Fig. 2 Generalized GPS user system configuration with separate delay estimating receiver and position-estimating functions. There are N satellite signals tracked by the parallel DLL ($2N$ signals if both L_1 and L_2 channels are used).

demodulators provide estimates of the pseudorange, carrier phase, and navigation data for each satellite and are usually implemented using digital processing. As discussed in Chapter 5, this volume, on the GPS satellite constellation, there may be as many as ten satellites in view at one time. Typically the number of parallel tracking DLL varies from 2 to 16, and it is possible to track all of the satellites in view at both L_1 and L_2 frequencies. Generally, at least five satellites are tracked as a minimum, either in parallel or in time sequence. At the present state of the art, an L_1 10-channel receiver with 10 parallel DLLs can be implemented in digital form on one CMOS chip. Semiconductor technology is increasing significantly every 18 months, and the processing power available for GPS receivers is expanding in the same manner.

The receiver system with its parallel signal DLL and computer processors must carry out the operations of satellite selection, signal search, tracking, and data demodulation, as shown in Table 1.

This chapter concentrates on the signal-tracking task. The other tasks are described in detail in Chapter 8, this volume. The parallel pseudoranges, navigation data, and, in the more sophisticated receivers, the carrier phase measurements are fed to the navigation data processor where the position of each satellite is calculated, and pseudorange and clock corrections are made. As a first step in this operation, the pseudorange and carrier phase measurements are corrected for the various perturbations, including satellite clock errors, Earth rotation effects, ionosphere delay, troposphere delay, relativistic effects, and equipment delays.

The corrected pseudorange, phase, or accumulated phase (ADR) measurements along with other sensor data are then fed to the extended Kalman filter (EKF) or similar filter. The output of the EKF estimator provides position, velocity, and time estimates relative to the user antenna phase center (see Chapter 9, this volume.) There may or may not be information from additional sensors; e.g., altimeters, inertial measurement units (IMUs), or dead-reckoning instruments. If so, these measurements are also fed to the EKF. Data from carrier tracking and some of these sensors can also be used to aid the DLL tracking operation itself. As discussed in the companion volume, this Kalman filter estimate of user position can also be used in a differential mode with other GPS receivers where at least one receiver is at a known reference point for geodetic sensing, more accurate airborne or shipborne navigation, or in common-view mode for precision differential time transfer. The user position is usually computed in Earth-centered, Earth-fixed (ECEF) coordinates and are then transferred by appropriate geodetic

Table 1 Simplified sequence of operations in a GPS receiver system

1) Select the satellites to be tracked among those in view. Approximate satellite position can be determined using the Almanac, and the selection criteria can be based on GDOP.

2) Search and acquire each of the GPS satellite signals selected.

3) Recover navigation data for each satellite.

4) Track the satellites under whatever conditions of user dynamics are present and measure pseudorange and range-rate and/or ADR.

transformation to a desired local coordinate set or map display convenient to the user. As pointed out in a previous paragraph and discussed in more detail later, it is also possible to integrate the EKF with the DLL instead of performing these operations independently. The author has termed the integrated system the vector delay lock loop (VDLL) because it processes the signals in parallel as a vector operation. The VDLL can be interpreted as a further extension of the EKF.

C. Alternate Forms of Generalized Position Estimators

Figure 3 shows the received GPS signals from N satellites, each with delay $\tau_i(\boldsymbol{x})$ and signal amplitude a_i. The user position is $\boldsymbol{x}$. The C/A or P received signal is expressed as follows;

$$r(t, \boldsymbol{x}) = \sum_{i=1}^{N} a_i d_i(t) s_i[t - \tau_i(\mathbf{x})]\cos[\omega_o(t - \tau_i) + \phi_i] + n(t) \tag{1}$$

where $s_i(t)$ represents the radio frequency signal* transmitted by satellite i with amplitude a_i, delay τ_i, and phase ϕ_i, and d_i is the binary data modulation. Both the delay and phase vary with time in accord with user dynamics relative to the satellite i. The signal s_i can represent either the C/A or P codes at either L_1 or L_2 frequencies†. The noise $n(t)$ is assumed to be stationary AWGN.‡ The delay

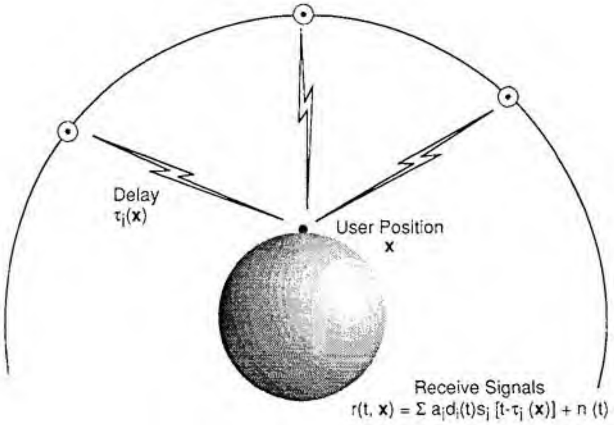

Delay
$\tau_i(x)$
User Position
x
Receive Signals
$r(t, x) = \Sigma\, a_i d_i(t) s_i\,[t-\tau_i\,(x)] + n\,(t)$

Fig. 3 Simplified representation of received GPS signals for a user at position x and path delays $\tau_i(x)$. The signals are received in the presence of white Gaussian noise $n(t)$.

*In reality, of course, the signal has both a C/A code and a P code in-phase quadrature. Both of these pseudonoise (PN) signals can be used simultaneously in delay estimation by a simple generalization of these results.

†More precisely, of course both C/A and P codes should be shown in phase quadrature, and both signals can be tracked in one receiver.

‡Although $n(t)$ is generally white Gaussian thermal noise, there clearly can be signal level fluctuations caused by multipath, attenuation caused by blockage from physical obstruction, and satellite-user motion/geometry. The delay $\tau_i(x)$ is also perturbed by atmospheric delay effects, selective availability, and multipath, which are separate colored noise effects discussed later.

τ_i and phase ϕ_i are functions of the user position vector $\boldsymbol{x}$. In general, the position vector $\boldsymbol{x}(t)$ is a function of time and is assumed to have some limited set of dynamics governed by a process model.

There are at least two general forms of position estimators that can be configured for the GPS system and many types within each of those forms. The most general estimator, shown in Fig. 4a, shows a single estimator that produces in one step an estimate of the user position vector $\boldsymbol{x}$. The second more restricted form of Fig. 4b first processes the received signal so as to estimate each of the satellite pseudoranges τ_i (and range-rate, etc.) and then generates an estimate of position $\hat{\boldsymbol{x}}$ based on the $\hat{\tau}_i$. Each of the estimates of τ_i are performed completely independently in this second form.

Assume for the moment that the user position is an unknown constant vector and that the received signal samples at discrete times t_k are $r(t_k,\boldsymbol{x})$ and are represented over the time interval $t_1, t_2, \ldots\ t_k$ as the following received vector:

$$\boldsymbol{r}(\boldsymbol{x}) = [r(t_1, \boldsymbol{x}), r(t_2, \boldsymbol{x}), \ldots, r(t_K, \boldsymbol{x})] \tag{2}$$

Then the estimate in Fig. one 4a can be represented as follows:

$$\hat{\boldsymbol{x}}_a = F_a[\boldsymbol{r}(\boldsymbol{x})] \tag{3}$$

The estimate of Fig. 4b is the following more restricted type of estimate;

$$\hat{\boldsymbol{x}}_b = F_b[\boldsymbol{r}(\boldsymbol{x})] = F_{bx}\{\hat{\tau}[\boldsymbol{r}(x)]\} \tag{4}$$

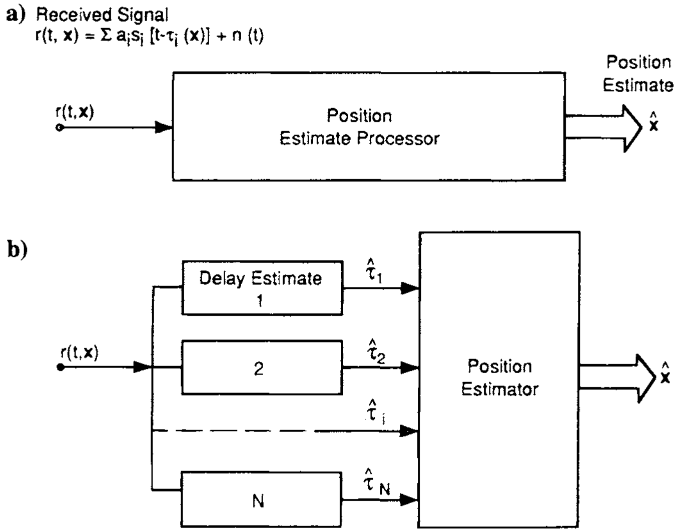


Fig. 4 Two forms of generalized GPS position estimate processors. The first form of processor a) estimates the position directly without an independent intermediate delay estimate. This estimator may also produce an estimate of delay, because with GPS a delay estimate is needed for recovery of the navigation data. The second form b) first estimates the delay using independent parallel estimators for each τ_i and then estimates position as a separate process.

where each of these τ_i delay (pseudorange) estimates in (b) are made independently. That is, no use is made of the fact that the delays τ_i may be correlated by the geometry of the transmission paths. Most present GPS receivers fall into this latter more restricted class.

An obvious potential disadvantage of the two-step approach of Fig. 4b can be illustrated by a hypothetical example where there are a large number of equal power satellite signals, say $N = 100$, and only one coordinate, a scalar x, to be estimated. Further assume that the satellite path delays are simple offsets of one another, namely, $\tau_i(x) = x + C_i$, where the C_i are known delay offsets. This signal model obviously represents an extreme example of an overdetermined estimation problem. Clearly, an estimate of x, based on the totality of 100 equal power received signals can, in general, produce a better estimate by making use of the linear relationship of the 100 different τ_i and the full power of the 100 signals than would an estimate based on processing each individual signal independently, and then combining the 100 independent estimates of delay. For example, a receiver could offset each of the signals by the known delays C_i add the signals coherently,* and estimate the delay using the composite signal that now has a signal-to-noise ratio 100 times as large. We can envision a situation where the signal-to-noise ratio of each individual signal is too small to process independently (i.e., below threshold); whereas, the delay in the composite signal is easily measured.

Section II of this chapter discusses receivers of the form of Fig. 4b wherein each delay estimate is made independently. Quasioptimum forms of the receivers (scalar delay lock loops) are discussed and related to more conventional delay lock loop tracking systems. The chapter concludes (Sec. III) with a discussion of quasioptimum forms of receivers of the more general form of Fig. 4a, which the author has defined as the vector delay lock loop (VDLL).[4]

D. Maximum Likelihood Estimates of Delay and Position

As shown in Appendices A and B, the DLL is a quasioptimal iterative form of two different statistically optimum delay estimators, the maximum likelihood estimator and the least mean square error estimator. In this subsection, we review a class of estimates termed the maximum likelihood estimate.[5-10] The parameter to be estimated can represent either delay or user position, is assumed to be constant over some time interval of interest, and slowly varies from interval to interval. When cast in an iterative approximation form, these estimators take the configuration of the DLL or the VDLL.

Assume that we make independent observations of a set of scalar random variables r_i, each of which depends on an unknown vector parameter $\boldsymbol{x}$ that we want to estimate. Define $\boldsymbol{r}$ as a vector representing a set of K random variables r_i, where the vector has a probability density $p(\boldsymbol{r})$. The conditional probability density of $\boldsymbol{r}$ is conditioned on a certain value for the unknown vector parameter $\boldsymbol{x}$, and is defined as $p(\boldsymbol{r}|\boldsymbol{x})$. The conditional density, $p(\boldsymbol{r}|\boldsymbol{x})$, is termed the likelihood function when the variable $\boldsymbol{x}$ is assumed to be the unknown random variable (rather than the opposite). In particular, suppose that we make K independent

*The noise also adds for each signal, but not coherently.

observations of the random variables r_i at discrete times t_k to form an observation vector $\boldsymbol{r} = (r_1, r_2, \ldots r_K)$, where the observations include additive Gaussian noise. The unknown $\boldsymbol{x}$ is assumed to be constant over this K sample interval. We define the maximum likelihood estimate as that estimate $\hat{\boldsymbol{x}}$, which maximizes $p(\boldsymbol{r}|\boldsymbol{x})$; namely,* the following:

$$p(\mathbf{r}\,|\,\hat{\mathbf{x}}) \geq p(\mathbf{r}\,|\,\mathbf{x}_o) \tag{5}$$

where $\boldsymbol{x}_o$ is any other estimate of $\boldsymbol{x}$.

Appendix A derives the maximum likelihood estimates for (1) delay $\tau(t_k)$; (2) a scalar position variable $x(t_k)$ where delay is a nonlinear function of x; and (3) a vector position variable $\boldsymbol{x}(t_k)$, where all signals are received in the presence of white Gaussian noise.

Figure 5 shows the block diagrams of the estimators (delay lock loops) that provide iterative closed-loop approximations to maximum likelihood for the two forms of estimators of Fig. 4 with time compared to the noise effects. Thus, the effective bandwidth of the closed-loop tracking filters can be small because it need only track the dynamics of user motion rather than the wide bandwidth of the signal itself. The estimator of Fig. 5a is configured as a closed loop operation in a form similar to what has been called the delay lock loop in earlier papers by the author and has the same form as the delay estimator of Fig. 4b.† The estimator of Fig. 5c is in a form that corresponds to Fig. 4a and is termed a vector delay lock loop.

The reader is referred to Appendix A for details. Suffice it to say here that each of these forms of delay lock loops serves to track the variable to be estimated by generating a correction term that is directly proportional to the error in the estimate. Furthermore, each tracking loop begins by forming the product of the received signal with a differentiated version of the signal component; namely, $s_i'(t)$.

A third closely related estimate is the least mean square error estimate for the AWGN channel. As shown in Appendix B, the least-mean-square estimate of a parameter for signals received is produced by a receiver that can also be configured in an iterative closed-loop form very similar to that for the maximum likelihood estimate when certain linearizing assumptions are made in both estimators. Thus, both estimators configured in iterative form are versions of delay lock loops.

E. Overall Perspective on GPS Receiver Noise Performance

Although this chapter discusses only the fundamentals of GPS signal tracking, it is important to place the noise performance that is discussed here in perspective with the noise effects on the other operations that must be performed in the GPS

*Another type of optimal estimate is the maximum a posteriori estimator (MAP),[11] which maximizes $p(\boldsymbol{x}|\boldsymbol{r})$; i.e., the most probable value of $\boldsymbol{x}$ given the observation vector $\boldsymbol{r}$. The MAP can be written $p(\boldsymbol{x}|\boldsymbol{r}) = \dfrac{p(\boldsymbol{r}|\boldsymbol{x})\,p(\boldsymbol{x})}{\int p(\boldsymbol{r}|\boldsymbol{x})\,p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}$. Thus, use of the MAP estimate requires knowledge of the a priori probability density $p(\boldsymbol{x})$ as well as $p(\boldsymbol{r}|\boldsymbol{x})$. However, in many problems, the a priori probability $p(\boldsymbol{x})$ is assumed to be unknown.

†For a pure sinusoidal signal, Viterbi[12] showed that the MAP estimator for a Gaussian phase variable can be configured in the form of a phase locked loop. For a pure sinusoidal signal, the delay lock loop reduces to a phase locked loop. See also Ref. 10.

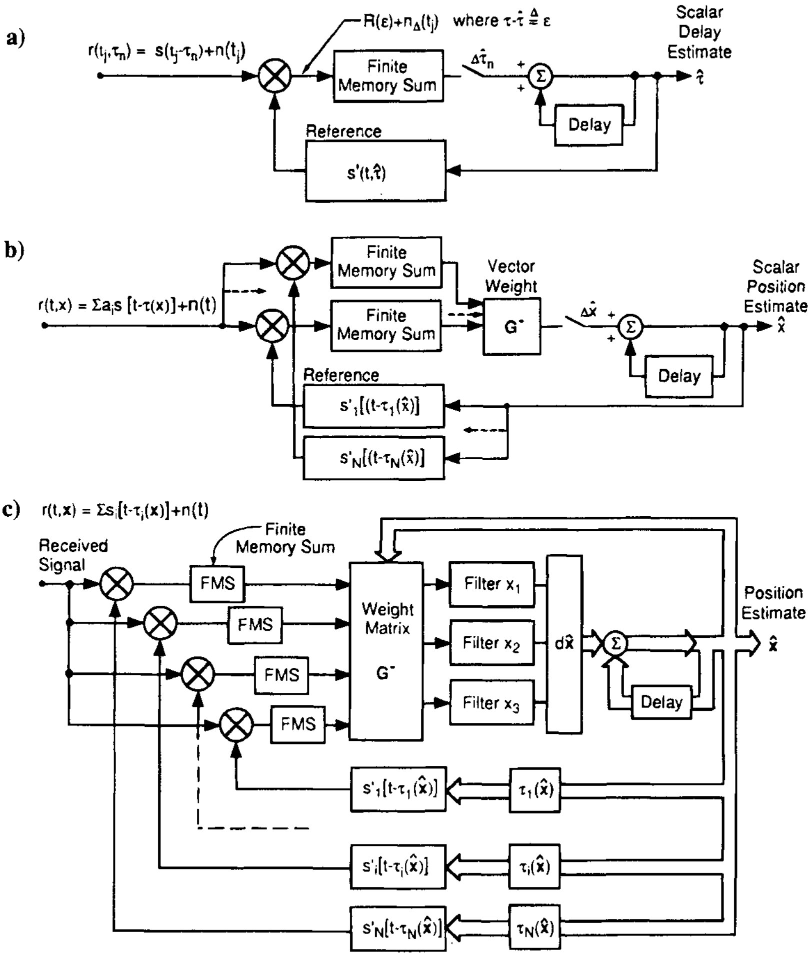


Fig. 5 These various forms of the delay-lock tracking loop provide estimates are iterative closed-loop approximations to the maximum likelihood estimates of τ or x provided that the changes in x occur in steps every K samples and are sufficiently small. Also assume that the loop begins with an initial estimate $\hat{\tau}$ or $\hat{x}$, which is close to the true value of τ or x. Three different generalizations of the delay lock estimator are shown for various levels of complexity of the estimation task. In part a, only a single received signal is present, and the parameter to be estimated is a scalar τ. In part b, multiple signals are received, and the quantity estimated is a scalar position variable x. In part c, multiple signals are received, and the parameter to be estimated is a position vector $x = (x_1,x_2,x_3, \dots x_m)$. The G^- matrix is the generalized inverse that gives the minimum mean square error estimate of dx given dτ. Different loop filters process each of the position variable components.

receiver; namely, GPS signal search and acquisition, and navigation data demodulation.

It is also important to illustrate some contrasts in the multiple access performance of GPS used for signal tracking with the multiple access performance of code division multiple access (CDMA) used for communications. The primary differences are twofold. First, the GPS signals are transmitted with only moderate power and are generally received with low gain, nearly omnidirectional antennas. Consequently the received radio frequency signal-to-noise ratios that result are low. Thus multiple access and self-noise are generally (although not always) of secondary importance to thermal noise effects. Second, with GPS, the primary objective is to estimate the user position and velocity, rather than transmit and receive a multiplicity of communications signals. Thus, as the number of satellites increases, the information sought does not increase significantly.* Contrast that objective with that of a CDMA communications network wherein each user signal carries with it digitized information at a substantial datarate, the transmission of which is the key objective of the communications system. In a CDMA system, maximum channel capacity is achieved when the signal-to-thermal noise ratio is large, thus multiple access noise effects dominate and limit the number of signals that can be transmitted (because each signal carries additional, rather than redundant, information).

Search and acquisition of the GPS signal is discussed in the next chapter. However, since this chapter on signal tracking assumes that the initial value of delay error has been reduced to a small value, it is well to spend a moment to make plausible that that assumption is realistic. The required C/N_o must be reasonable.

GPS signal search and acquisition is often accomplished by sequencing a reference C/A code over each of the 1023 chips of the C/A code in fractional chip (often in 1/2 chip) steps, in each of several Doppler frequency offset increments. This time-frequency search is completed when the signal component is detected by a noncoherent square law detector. The noncoherent detector operates using a relatively narrow IF bandwidth W, and acquisition occurs when the detector produces an output that exceeds some threshold level. The threshold level is set, in turn, to produce some acceptable level of false alarms, perhaps on the order of 5 in 10^3. If threshold is exceeded several times; e.g., three times out of five, "lock" is declared. If each time-frequency cell is examined for T_r s; i.e., the search rate is $S_r = 1/T_r$. For a false alarm probability $P_{FA} = 5 \cdot 10^{-3}$, and a probability of detection $P_d = 0.9$, the required ratio of carrier power $C = P_s$ to noise density, N_o (one-sided) is approximately $C/N_oS_r \cong 22$ or 13.4 dB if the IF signal-to-noise ratio is unity or more.[3,13] Thus if† $C/N_o = 33$ dB-Hz, and the IF bandwidth is 2000 Hz, then the maximum search rate is $S_r < 90$ or 45 chips/s if 1/2 chip steps are used. If the maximum residual Doppler phase oscillator drift is ± 1 kHz, then only one IF channel of 2 kHz must be searched.

*There is, of course, more navigation data to be recovered but these data are at a very low datarate of 50 bps.

†Typical GPS C/N_o are larger; $C/N_o > 40$ dB-Hz. The rule of thumb $C/N_oS_R = 22$ for signal search can also be written $E_R/N_o = 22$ where $E_R = C/S_R$ is the energy per search time interval $1/S_R$. Note the similarity in form with the bit error rate constraint $E_b/N_o = 10$ where E_b is the energy per bit.

Thus search and acquisition do not pose any severe constraints on C/N_o for these search rates. Search and acquisition are discussed more fully in the next chapter.

The navigation data demodulation places a similar requirement on C/N_o. Coherent demodulation of the binary phase-shift keyed (BPSK) waveform produces a theoretical bit error probability $P_E = \mathrm{erfc}\sqrt{2E_b/N_o} = \mathrm{erfc}\sqrt{2C/N_of_d}$ where E_b is the energy per bit $= P_sT_d = CT_d$, where $T_d = 1/f_d$, and f_d is the 50 bps datarate. If $C/N_of_d = 10$, then the output error probability is approximately 10^{-5}, which is quite satisfactory for good receiver performance. This ratio corresponds to a $C/N_o = 500$ or 27 dB, which is less than the 33 dB-Hz just cited in the search and acquisition example.

Thus, we can proceed with the discussion of tracking knowing that the C/N_o requirement for acquisition in a reasonable time is consistant with available GPS signal levels (see Chapter 3, this volume), and data can be demodulated at sufficiently low error rate.

F. Interaction of Signal Tracking and Navigation Data Demodulation

As described above, the complete GPS receiver has two closely related and interacting tasks; the receiver must track the delay of each of the received signals and must also coherently (or differentially coherently) detect the 50 bps navigation databit stream, as shown in Fig. 6. Each of these tasks can support the other. For example, the extraction of the navigation data relies on a reasonably accurate delay estimate of each signal so that a reference waveform properly aligned in time; i.e., punctual, can be used to remove the GPS spread spectrum PN code and recover the BPSK signal. The BPSK signal itself can be coherently demodulated by recovering an estimate carrier phase $\phi_i(t)$ and then extracting the navigation data $d_i(t + \tau)$. The recovered carrier phase and navigation data estimates can, in turn, be used as a coherent reference to eliminate the carrier and data in the

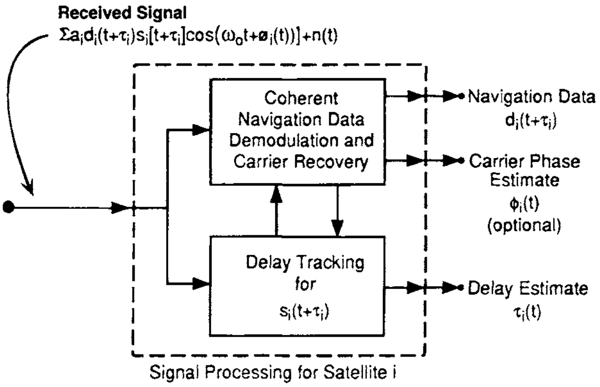

Received Signal
$\Sigma a_i d_i(t+\tau_i)s_i[t+\tau_i]\cos(\omega_0 t+\phi_i(t))]+n(t)$
Coherent Navigation Data Demodulation and Carrier Recovery
Navigation Data
$d_i(t+\tau_i)$
Carrier Phase Estimate
$\phi_i(t)$
(optional)
Delay Tracking for $s_i(t+\tau_i)$
Delay Estimate
$\tau_i(t)$
Signal Processing for Satellite i

Fig. 6 Two separable, but interacting, signal processing tasks in the GPS receiver, navigation data demodulation, and signal delay tracking. The outputs of each of these tasks can aid the operation of the other.

received signal to permit coherent tracking. Alternatively, this carrier frequency estimate can be used as an aid in removing Doppler shift from the received signal prior to the delay tracking operation, or as an aid in code tracking.

II. Delay Lock Loop Receivers for GPS Signal Tracking

This section describes several alternate methods for tracking the GPS signal. Both code tracking and, to a lesser extent, carrier tracking are described. The signal is generally assumed to be received in the presence of stationary additive white Gaussian noise (AWGN). In this chapter, the signals $s_i(t)$ are always assumed to be bandlimited finite rise time signals with a continuous autocorrelation function $R(\epsilon)$ and continuous, finite, differentiated autocorrelation function $R'(\epsilon) \;\Delta\; \partial\; R(\epsilon)/\partial\epsilon$. The discussion begins with the quasioptimal delay lock loop that employs a differentiated signal as the reference.[1,3] This delay lock loop is then related to the early-late gate delay lock loop.[2,3,14] It is shown that for a signal represented by a PN trapezoidal pulse sequence with finite risetime, the two forms are identical. Although some of the discussion of tracking systems is expressed in terms of continuous time systems, the reader should be aware that implementations are generally performed in discrete time digital operations. Thus, read $s(t_k)$ for $s(t)$ where the samples are taken at discrete times t_k with independent samples taken at the Nyquist rate.

A. Coherent Delay Lock Tracking of Bandlimited Pseudonoise Sequences

Consider a coherent tracking receiver as shown in Fig. 7 where for the moment we focus on a single received PN signal of power $P_s = A^2/2$ plus bandpass AWGN of spectral density N_o (one-sided).* The coherent receiver downconverts the signal and removes the data modulation using information and control signals from a yet to be described coherent demodulation channel (see Sec. I.F and II.C). The output of the coherent downconverter is then passed through a low-pass

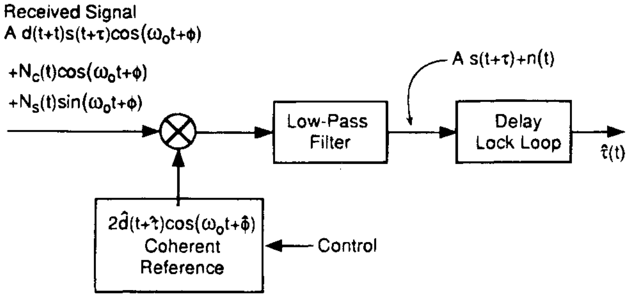


Fig. 7 General form of the coherent delay lock loop receiver system.

*Throughout much of this section, the functions are often expressed in continuous time t. The reader should be aware that these results are easily translated to complex discrete time functions at sample time t_k sampled at the Nyquist rate.

filter that maintains the signal spectrum but removes the components at frequency $2f_o$. The output of the coherent downconverter, assuming sufficiently accurate estimates of carrier phase ϕ and data d, is then a pure PN signal plus white Gaussian noise $n(t)$. It should be pointed out that coherent tracking loops of this form have definite practical limitations in that a carrier tracking cycle slip can cause the code tracking loop to lose lock. Nonetheless, the coherent tracking loop serves a useful purpose as a bound on performance and as an introduction to the other forms of the DLL.

The output of the multiplier low-pass filter combination in Fig. 7 for perfect data demodulation and carrier phase recover, namely, $\hat{d}(t) = d(t)$ and $\phi = \hat{\phi}$, is then $A\ s(t + \tau) + N_c(t)$. The radio frequency signal power is $P_s = (A^2/2)$ for constant envelope signals and somewhat less than that for finite rise time trapezoidal waveforms. The ratio of the signal power in the multiplier output $A^2 = 2P_s$ to the noise density* $2N_o$ of the $n(t) = N_c(t)$ term is then $A^2/2N_o = P_s/N_o$, the same as at the radio frequency input. The baseband pseudonoise PN signal is represented by the following:

$$s(t) = \sum_i P_i p(t - iT)$$

where P_i is a random or pseudorandom binary sequence of numbers ±1, and $p(t)$ is a bandlimited finite rise time pulse waveform of approximate pulse width T. The PN clock rate is $f_c = 1/T$.

The original paper on the "Delay-Lock Discriminator"[1] pointed out that the optimal tracking system for tracking the delay of a signal $s(t + \tau)$ can be approximated by a tracking loop that first multiplies the received signal plus noise by the differentiated signal, as shown in Fig. 8. The true delay τ and

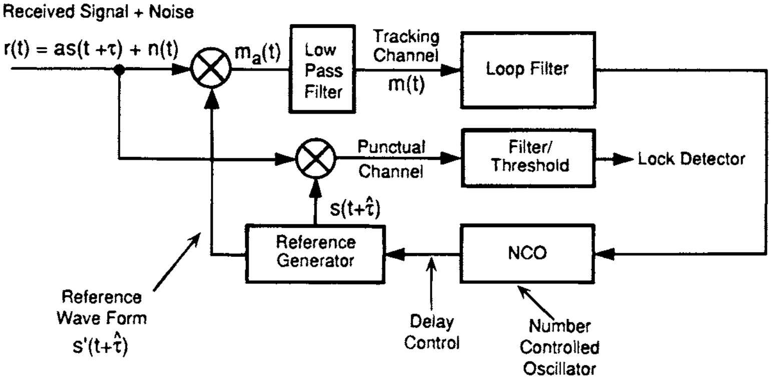


Fig. 8 Delay lock tracking of signal $s(t + \tau)$ where τ varies with time. The punctual channel reference $s(t + \hat{\tau})$ is used to detect that the receiver is locked-on. The bandwidth of the closed-loop tracking must match the dynamics of the delay variation.

*The power of the radio frequency finite bandwidth AWGN, $N_c(t)\cos\omega_o(t) + N_s(t)\cos\omega_o(t)$ with center frequency ω_0 is $P_n = (1/2)E[N_c^2(t) + N_s^2(t)] = E[N_c^2(t)]$. If the one-sided spectral density of the radio frequency noise is N_o, then the one-sided spectral density of $N_c(t)$ is $2N_o$, noting that $N_c(t)$ has half the bandwidth of the radio frequency noise.

the delay estimate $\hat{\tau}$ are assumed to be sufficiently close that first-order linear approximations are accurate; i.e., the delay error $\epsilon \; \Delta \; \tau - \hat{\tau}$ is sufficiently small relative to the width of the autocorrelation function. The received signal, $s(t + \tau)$, can then be expanded in a Taylor series referenced to the signal with a delay estimate $\hat{\tau}$ as follows:

$$s(t + \tau) \cong s(t + \hat{\tau}) + \epsilon s'(t + \hat{\tau}) + \frac{\epsilon^2}{2} s''(t + \hat{\tau}) +, \ldots, \text{ for small } \epsilon \qquad (6)$$

where the partial derivative of the signal waveform with respect to time is $s' \Delta \; (\partial/\partial t) \; s$, etc.

The output of the upper multiplier, the tracking channel, in Fig. 8 is $m_a(t)$, which can be expanded using Eq. (6) as follows:

$$m_a(t) = s'(t + \hat{\tau})[As(t + \tau) + n(t)] \qquad (7a)$$

$$m_a(t) \cong As(t + \hat{\tau})s'(t + \hat{\tau}) + A\epsilon(t)s'(t + \hat{\tau})^2$$
$$+ (A\epsilon^2(t)/2)s''(t + \hat{\tau})s(t + \hat{\tau}) +, \ldots, + n(t)s'(t + \hat{\tau}) \qquad (7b)$$

As before, $n(t)$ is assumed to be stationary AWGN is independent of $s(t)$. The signal is also assumed to be stationary. Define $R(\epsilon) = E[s(t + \epsilon)s(t)]$ as the autocorrelation function of the signal $s(t)$. Note that the expected value of the multiplier output is as follows:

$$E[m_a(t)] = E(s(t + \tau)s'(t + \hat{\tau})] = E[s(t + \hat{\tau} + \epsilon)s'(t + \hat{\tau})] = R'(\epsilon) = D(\epsilon) \qquad (8)$$

where $R'(\epsilon) = D(\epsilon)$ is the differentiated autocorrelation function where $R'(\epsilon) = (\partial/\partial\epsilon)R(\epsilon)$, and $D(\epsilon)$ is termed the delay lock loop discriminator characteristic. The slope of the discriminator characteristic at $\epsilon = 0$ is $D'(0)$ and defines the loop gain of the DLL. Define $P_d = E[s'(t)]^2 = -R_s''(0) = -D'(0)$ as the power of the differented signal,* and assume that the clock-rate f_c of the signal $s(t)$ is large compared to the closed-loop noise bandwidth† B_n of the linearized equivalent tracking loop, which is defined later. We can represent the product $s(t)s'(t + \epsilon)$ by its expected value plus a "self-noise" term‡; namely,

$$s(t + \tau) \; s'(t + \hat{\tau}) = D(\epsilon) + [s(t + \tau) \; s'(t + \hat{\tau}) - D(\epsilon)] = D(\epsilon) + n_{\text{sn}}(t)$$

where $n_{\text{sn}}(t)$ is a wide bandwidth self-noise term

$$n_{\text{sn}}(t) \; \Delta \; [s'(t + \hat{\tau})s(t + \tau) - D(\epsilon)],$$

*The slope of the discriminator characteristic is $D'(0) = R_s''(0)$ and is negative. Note that the dimension of $s'(t)$ is s^{-1} and the dimension of $(s')^2$, and hence, P_d is s^{-2}.

†If the closed-loop transfer function is $H(j\omega)$ then the closed-loop noise bandwidth is

$$B_n = \frac{1}{2\pi} \int_0^\infty |H(j\omega)|^2 \, d\omega / |H(0)|^2$$

‡Self-noise is defined as the difference between the expected value of the product $s(t + \tau)$ $s'(t + \hat{\tau})$ and its actual value. The self-noise is a broadband noise-like waveform.[3,4]

most of which is removed by the low-pass and tracking loop filters, and thus, can often be neglected.

The output of the upper multiplier $m_a(t)$ in Fig. 8 is next filtered with a low-pass filter. This filter has bandwidth that is small compared to f_c but large compared to the dynamics of the delay. Thus, it performs a limited amount of averaging. Hence, it is useful to represent the low-pass filtered multiplier output $m(t)$ of Eq. (7) as follows:

$$
\begin{aligned}
m(t) &= AD(\epsilon) + [n(t)s'(t + \hat{\tau}) + An_{\text{sn}}(t)|_{\text{low pass}}] = AD(\epsilon) + n_s(t) \\
m(t) &\cong -A\epsilon(t)P_d + [n(t)s'(t + \hat{\tau}) + An_{\text{sn}}(t)] \\
&\quad + (\epsilon^2, \epsilon^3 \text{ terms})|_{\text{low pass}} \\
&\cong -A\epsilon(t)P_d + n_s(t) \qquad \text{for small } \epsilon \text{ and } B_n << B
\end{aligned}
\tag{9}
$$

where $D(\epsilon) \approx D'(0)\epsilon = -P_d\epsilon$ for small ϵ, self-noise effects are neglected, and it is assumed that the ϵ^2 and higher-order ϵ terms are negligible for small ϵ and $n_s(t) = n(t)s'(t + \hat{\tau})|_{\text{low pass}}$. Thus, the last form of Eq. (9) represents the first two terms in the Taylors' series of Eqs. (6) and (7b). The first component of $m(t)$ in Eq. (9) represents the delay lock loop discriminator characteristic* $D(\epsilon)$.

Thus, for small ϵ, the output of the low-pass filter of Fig. 8 produces a correction term directly proportioned to delay error ϵ, and thereby, enables the loop to track the delay $\tau(t)$. To begin the tracking operation we must initialize (lock up) the DLL with a moderately accurate initial estimate of τ obtained by some search procedure (see Chapter 8, this volume). Note also that for a pure sine wave signal $s(t) = \sin \omega t$, the differentiated signal $s'(t) = \omega \cos \omega t$, and the delay lock loop for this special case simplifies to a conventional phase lock loop. In this case, the filtered product $s(t + \tau)s'(t + \hat{\tau}) = \omega \sin \omega\epsilon \cong \omega^2\epsilon$ for $\omega\epsilon << 1$ (neglecting the 2ω terms), and the discriminator characteristic is $D(\epsilon) = \omega \sin (\omega\epsilon)$. Digital phase lock loops are employed in the carrier-tracking operation.

1. Trapezoidal Pseudonoise Waveform

It is useful to approximate the finite rise time PN waveform by a symmetrical trapezoidal waveform† $s(t)$ of zero mean (equiprobable zero and ones) and rise time δT, as shown in Fig. 9. This finite rise time PN waveform is a more realistic representation of the GPS signal than the ideal rectangular shape. Thus, the DLL reference waveform $s'(t)$ is a ternary waveform that is a sequence of narrow pseudorandom pulses with pulse width equal to the rise time δT. It can be seen that the synchronized product $s(t)s'(t)$ with zero delay offset has zero long-term average ($D(0) = 0$), as we would expect. For a finite rise time pseudorandom

*Note that with this definition of $D(\epsilon)$, the slope at $\epsilon = 0$ is negative, and we must put a sign reversal in the loop in order to track the delay variable.

†To be consistent with the assumption of a bandlimited waveform with a continuous derivative, assume that the corners of the trapezoid are slightly rounded.

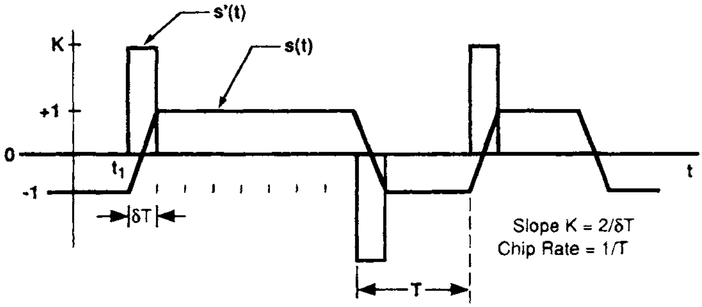


Fig. 9 Plot of PN sequence with a trapezoidal waveform and finite rise time $\delta T < T/2$. The differentiated waveform $s'(t)$ is also shown. The peak value of $s'(t)$ is $K = 2/\delta T$. Where there is no transition $s'(t) = 0$. Because there are transitions half the time, the duty factor of $s'(t)$ is $\delta T/2T$.

sequence of chip duration T, the multiplier output integrated over one chip interval T with zero delay error also is zero; namely

$$\int_{NT}^{NT+T} s(t)s'(t)\mathrm{d}t = 0$$

Note that for the trapezoidal waveform, the only portion of the waveform useful for time delay measurement is during the transient rise time δT of the waveform, which occurs at the beginning of the chips 50% of the time. This characteristic is in marked contrast to a communication channel where essentially all of the waveform is useful. Thus, the noise is time-gated to a duty factor of $\delta T/2T$ without significant loss of tracking information.

Thus, the multiplier output of Eq. (9) can be approximated as follows:

$$\begin{aligned} m(t) &\cong AD[\epsilon(t)] + n(t)s'(t)|_{\text{low pass}} \\ &\cong -AP_d\epsilon(t) + n(t)s'(t)|_{\text{low pass}} \qquad \text{for small } \epsilon \end{aligned} \tag{10}$$

where the multiplier output noise is time-gated by $s'(t)$, which has a duty factor* of $\delta T/2T$. The discriminator function for this trapezoidal PN sequence is quasilinear $D(\epsilon) \cong -[2\epsilon/\delta T - (\epsilon/\delta T)^2]$ for $|\epsilon| < \delta T/2$. Figure 10 shows a plot of example trapezoidal pulse waveforms, their autocorrelation functions $R(\epsilon)$, and differentiated autocorrelation functions $R'(\epsilon) = D(\epsilon)$ for unit amplitude triangular and trapezoidal pulse waveforms. The power in the differentiated signal is as follows:

$$-D'(0) = P_d = E[s'(t)^2] = K^2\frac{\delta T}{2T} = \left(\frac{2}{\delta T}\right)^2\frac{\delta T}{2T} = \frac{2}{T(\delta T)} \tag{11}$$

where $D'(0) = -P_d$, and $K = 2/\delta T$ is the slope of $s(t)$ at a transition, and it is assumed that the PN signal has transitions one-half the time. The rise time δT is inversely related to the one-sided 3-dB bandwidth of the baseband wave-

*The duty factor of each pulse is $\delta T/T$, but its pulses occur only in half of the chip intervals.

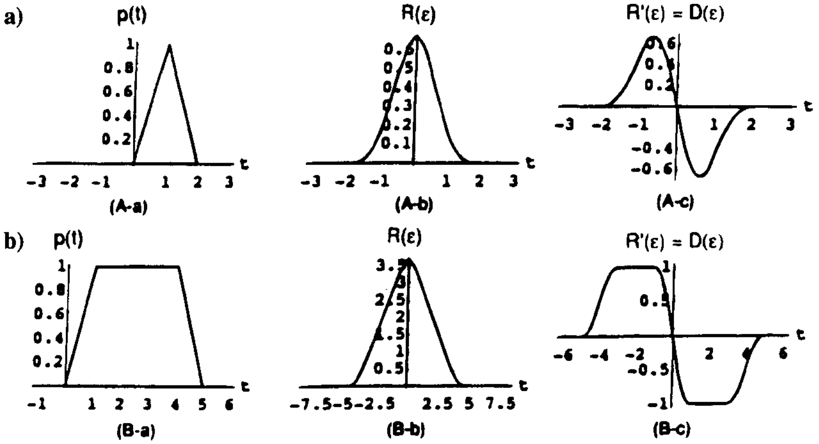


Fig. 10 Triangular- and trapezoidal-shaped pulses with the same normalized $\delta T = 1$, their autocorrelation functions and the quasioptimal delay lock discriminator functions $D(\epsilon) = R'(\epsilon)$. In part a, the PN sequence chip rate is normalized to 1.0, and the rise time of the PN pulse is also 1.0 leading to a triangular waveform shown in A-a with $\delta T/T = 1.0$. The autocorrelation function $R(\epsilon)$ is shown in A-b, and the differentiated waveform or quasioptimal delay lock loop discriminator characteristic $R'(\epsilon)$ is shown in A-c. Figure 10b shows the trapezoidal waveform B-a, autocorrelation function B-b, and discriminator characteristic $R'(\epsilon)$) in B-c, for a PN pulse where the rise time $\delta T = 1.0$ is 25% of the pulse width $T = 4.0$; i.e., $\delta T/T = 0.25$.

shaping filter B by the approximation* $\delta T \cong 0.44/B$. If, for example, the bandwidth $B = 1/T$, then $\delta T = 0.44T$. If, on the other hand, the bandwidth is considerably larger and $B = 10/\mathrm{T}$, as is approximately the situation for the transmitted C/A code on the GPS L_1 channel, then $\delta T = 0.044T$. Thus, the pulse width of the quasioptimum reference waveform is very narrow for short rise-time pulses. Even for the P code, the rise time is significantly less than T.

2. *Delay Lock Loop Discriminator Curve*

Assume again that the signal waveform is the trapezoidal shaped PN sequence. The reference waveform is then a PN sequence of finite width pair of rectangular pulses shown in Fig. 11a where $T = 4$, $\delta T = 1$. The coherent delay lock loop with a differentiated reference has a discriminator curve $D(\epsilon) = R'(\epsilon)$, as shown in Fig. 11b (and also Fig. 10B-c).

Thus, for small values of $\delta T/T$, this delay lock loop operates approximately as a piecewise-linear "bang-bang" servo system, and the discriminator curve

*The trapezoidal waveform can result by filtering a rectangular shaped PN sequence with a finite memory integrator filter with impulse response $h(t) = 1/\delta T$ for $0 \leq t \leq \delta T$, and $h(t) = 0$ otherwise. This filter has a frequency transfer function $|H(j\omega)| = \sin(\pi f \delta T)/\pi f \delta T$. This signal spectrum has a first zero in the frequency response at $f = 1/\delta T$. At frequency $f = 0.44/\delta T$, the spectral response is down 2.97 dB. Thus, we define the bandwidth measure $B = 0.44/\delta T$. Clearly, the actual relationship between δT and bandwidth of the waveshaping filter is dependent on the exact signal waveform and the definitions of bandwidth and rise time.

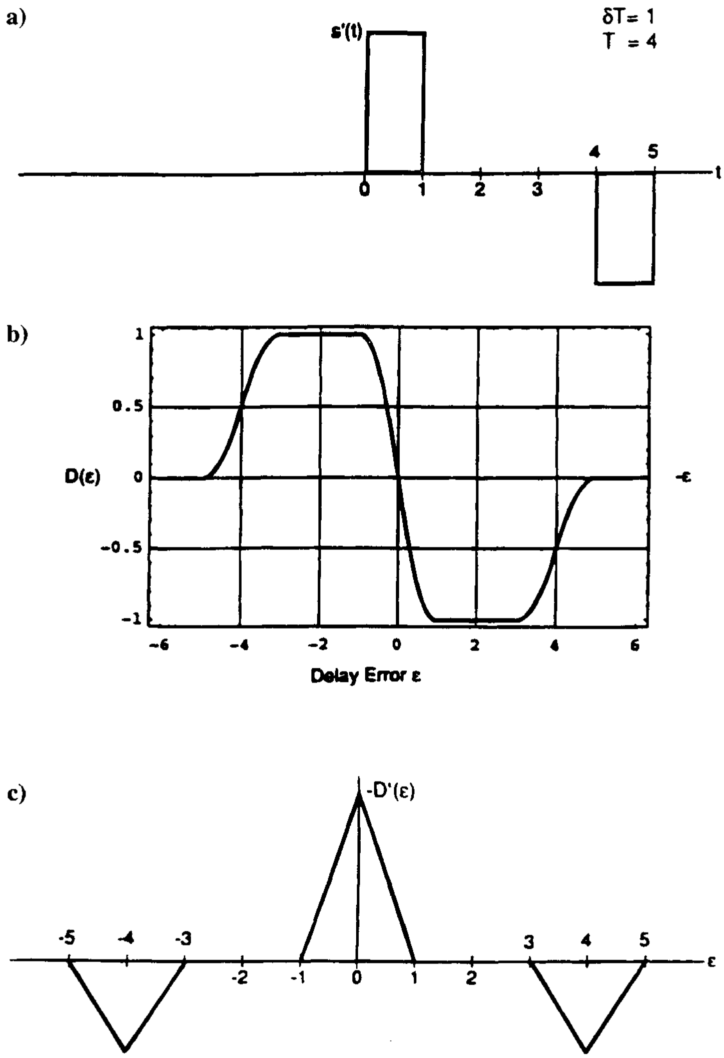


Fig. 11 Coherent delay lock loop discriminator curve $D(\epsilon)$ for a finite rise-time PN signal with a trapezoidal waveform for the optimal delay lock loop. The slope of $D(\epsilon)$,″ namely, $D'(\epsilon)$, is also shown. The differentiated reference s' is shown in part a. The discriminator curve $D(\epsilon)$ is shown in part b, and the derivative of the discriminator curve (loop gain) $D'(\epsilon)$ is shown in part c. The rise time $\delta T = 1$ and the pulse period $T = 4$ in this example. Thus, the normalized rise time is $\delta T/T = 0.25$. The slope $-D'(0) = 2/(T\delta T)$.

approximates a square wave "doublet." The slope $D'(\epsilon)$ of the discriminator function is as follows:

$$D'(\epsilon) = \frac{\partial}{\partial\epsilon} D(\epsilon) = -\frac{\partial}{\partial\epsilon}\int s(t+\epsilon)s'(t)\mathrm{d}t = -\int s'(t+\epsilon)s'(t)\mathrm{d}t \qquad (12)$$

and is zero for $|\epsilon/T| > 1 + \delta T/T$. If $s'(t)$ is a quasirectangular-shaped "doublet" of Fig. 11a then the slope of $D(\epsilon)$ is a triangular-shaped "triplet," as shown in Fig. 11c.

3. *Coherent Early-Late Gate Delay Lock Loops*

The earliest published delay lock loop specifically for binary PN signals used two reference signals, an early reference and a late reference signal, each binary PN signal with nearly rectangular shape[2] (see Fig. 12a).* Often the difference between the early signal and the late signal is set equal to a chip width T s. The low-pass filters in the two legs of Fig. 12a simply remove broadband noise and self-noise, as discussed in the previous section. Define the (nearly) zero rise-time binary PN reference waveform as $s_o(t)$. Clearly, the difference between the early and late binary reference signals shown in Fig. 13 is identical to the ternary signal at the bottom of Fig. 13 if the waveforms $s_o(t)$ are rectangular with zero rise-time.

The low-pass filtered output $m(t)$ for the early-late version of the DLL can be written from Fig. 12 as follows:

$$\begin{aligned} m(t) &= [As(t+\hat{\tau}) + n(t)][s_o(t+\hat{\tau}-\Delta/2) - s_o(t+\hat{\tau}+\Delta/2)]\,|_{\text{low pass}} \\ &\cong A[R_s(\epsilon-\Delta/2) - R_s(\epsilon+\Delta/2)] + n(t)s_\Delta(t+\hat{\tau})\,|_{\text{low pass}} = AD_\Delta(\epsilon) \\ &\quad + n(t)s_\Delta(t+\hat{\tau})\,|_{\text{low pass}} \\ &\cong AD'_\Delta(0)\epsilon + n(t)s_\Delta(t+\hat{\tau})\,|_{\text{low pass}} \end{aligned} \qquad (13)$$

where $R_s(\epsilon) = E[s(t+\epsilon)(s_o(t))]$ is the crosscorrelation between the reference $s_o(t)$ and the finite rise-time received waveform $s(t)$, $s_\Delta(t) = s_o(t+\Delta/2) - s_o(t-\Delta/2)$ and the DLL discriminator function is $D_\Delta(\epsilon) = R_s(\epsilon+\Delta/2) - R_s(\epsilon-\Delta/2)$. Note that this early-late DLL multiplier output, Eq. (13), is in exactly the same form as that for the quasioptimal DLL and is identical if $s_\Delta(t) = s'(t)$.

Thus, for the trapezoidal PN waveform, the $s'(t)$ ternary pulse sequence of Fig. 11 is exactly equivalent to the properly scaled difference between two time displaced zero rise-time PN sequences, as shown in Fig. 13. Thus, the differentiated delay lock loop of Fig. 12b is also equivalent to an early-late gate delay lock loop with delay offsets of $\pm\delta T/2$, as shown in Fig. 12a. In this chapter, the difference in delay between the early and late reference signals is defined as Δ.

For a signal with bandwidth $B = 10/T$ then if rise time is related to bandwidth by $\delta T = 0.44/B$, the quasioptimal delay differences $\Delta = \delta T = 0.044T$. This offset is significantly smaller than the commonly used delay difference $\Delta = T$.

*The original delay lock loop paper in 1961 defined the differentiated reference that gives an early-late spacing equal to the rise-time for a trapezoidal waveform. However, at that time, the state of the art made a longer delay spacing more practical.

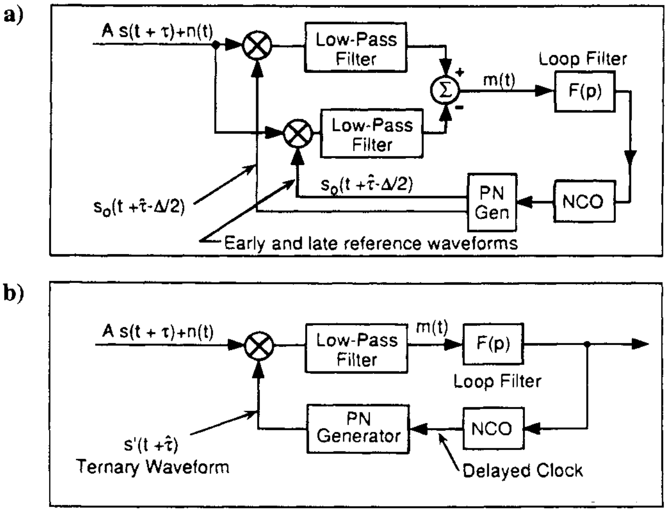


Fig. 12 The early-late delay lock loop with a delay difference between early and late reference signals of δ*T* (part a). For a trapezoidal PN waveform, this loop is mathematically equivalent to the loop shown in part b, the optimal delay lock loop. The early-late delay difference Δ is set at Δ = δ*T*, where δ*T* is the pulse rise-time. The NCO is a number-controlled oscillator used to control the clock phase of the PN generator. Part b shows the optimal delay lock loop for the trapezoidal or ternary waveforms of Fig. 9 utilizes a differentiated reference that is a ternary waveform.

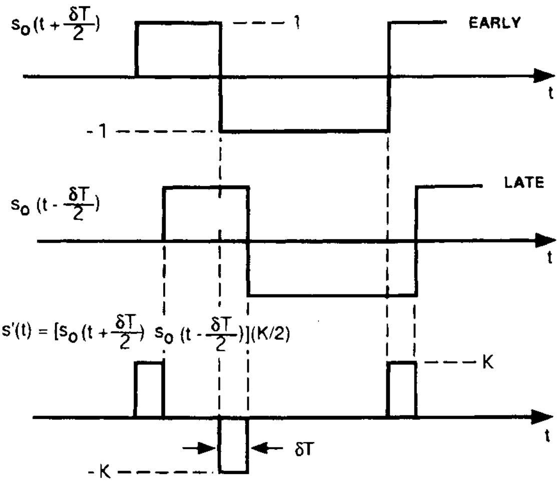


Fig. 13 The pulse sequence representing $s'(t)$ for a trapezoidal finite rise-time δ*T* waveform of Fig. 11 is equal to the difference between two zero rise-time reference PN sequences; namely, $s'(t) = [s_o(t + \Delta/2) - s_o(t - \Delta/2)](K/2)$ with only a scale factor *K*/2 difference if the delay difference is Δ = δ*T*.

The smaller delay offset gives a higher accuracy but a slightly smaller threshold acquisition range and a substantially smaller quasilinear region. As discussed later in the subsection on transient performance of the DLL, there are certain disadvantages if the early-late spacing Δ becomes too small.* Note, however, that if the rise-time of the PN signal δT is equal to the chip interval; i.e., $\delta T = T$, then the signal becomes the triangular wave of Fig. 10, and the quasioptimal early-late gate delay lock loop delay offset is the commonly used plus or minus half-chip setting.[2] Thus, the commonly used delay lock loop for an early-late spacing of $T/2$ is optimum only for the triangular wave PN signal.†

If the received signal has an ideal rectangular pulse shape, the discriminator characteristic $D(\epsilon)$ for the early-late reference signal with early–late offsets of $\pm T/2$; i.e., $\Delta = T$ is piecewise linear, as shown in Fig. 14, and has a normalized one-sided width $(\epsilon/T) = 1.5$, and thus, gives a nearly 50% wider acquisition range than the DLL with narrow correlator spacing for $\delta T/T << 1$. However, as shown later, the noise performance of the DLL that uses an approximation to the differentiated reference, a PN sequence of narrow pulses or equivalently a narrow early-late spacing close to the rise-time of δT, gives better noise performance than that for the early-late gate DLL with the wide early-late spacing, $\Delta = T$.

It is informative to compare the noise performance of two different delay lock loop discriminator characteristics that can be used for the same received trapezoidal pulse waveform of Fig. 9. Consider a pulse of width $T = 4$ for this waveform with a rise-time $\delta T = 1$ for a ratio $\delta T/T = 0.25$. The two different reference waveforms are shown in Fig. 15. Figure 16 shows the discriminator characteristic for the quasioptimal reference waveform $s'(t)$ as the solid curve where its discriminator characteristic $D(\epsilon) = R'(\epsilon)$. The dashed curve is the discriminator characteristic for an early-late gate delay lock loop with an early-late separation of $\Delta = T$. Note that both discriminator curves have the same slope at $\epsilon = 0$. However, the effective noise spectral density is four times as large for the loop with the $\Delta = T$ separation because the effective noise is time gated by the duty factor. Thus, the noise performance of the quasioptimal loop is superior for this trapezoidal signal for small delay error.

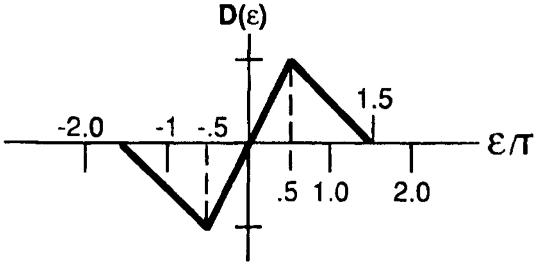


Fig. 14 Delay lock loop discriminator curves for offset reference signals $s(t \pm T/2)$; i.e., $\Delta = 1$, and an ideal rectangular received PN signal.

*During acquisition, we may have to increase the loop gain to improve the pull-in performance.

†Note that there are more recent papers that discuss the use of narrower time separations between the early and late gate signals. See also the next chapter.[15,16]

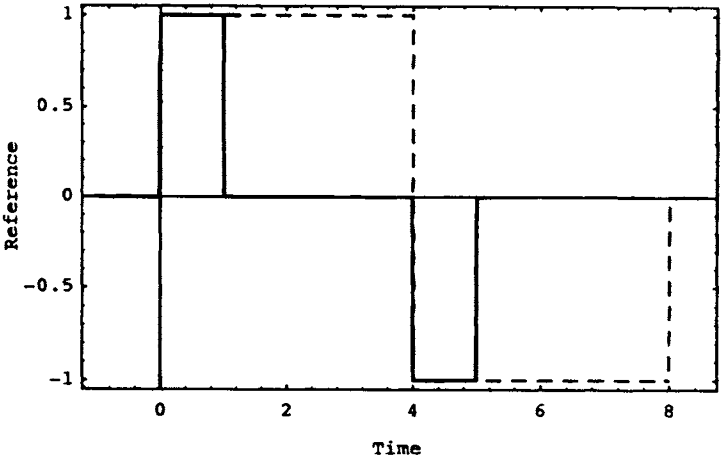


Fig. 15 Quasioptimal reference waveform $s'(t)$ for the trapezoidal shaped PN pulse waveform, $\Delta = \delta T = T/4$ (solid curve), and a reference with an early-late separation of T. The received Trapezoidal pulse waveform has a pulse duration of $T = 4$ and a $\delta T = 1$ rise-time. The dashed curve shows the commonly used early-late reference waveform with an early-late separation of $\Delta = T$ s.

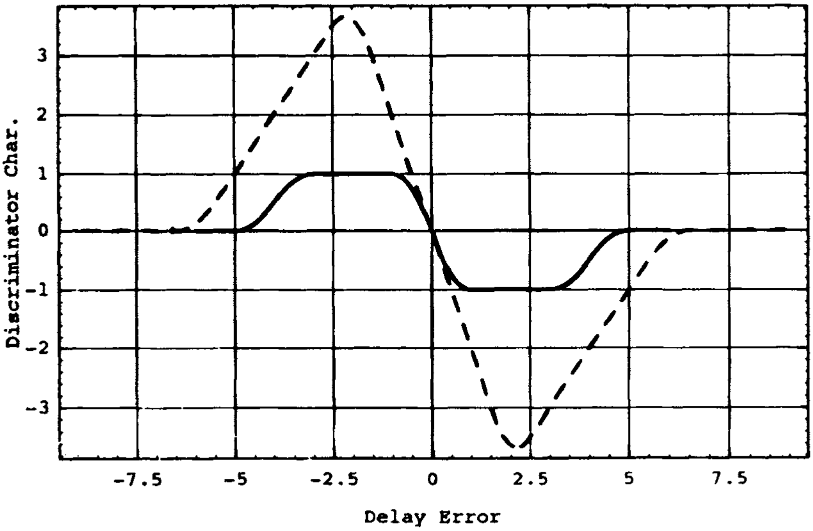


Fig. 16 Quasioptimal and early-late gate discriminator characteristics for the trapezoidal waveform of Fig. 10b with $T = 4$ and $\delta T = 1$ for a rise-time ratio $\delta T/T = 0.25$. The solid curve, quasioptimal, is obtained using the differentiated reference $s'(t)$. The dashed curve is obtained using an early-late gate rectangular reference with an early-late separation of T. Note that the slope of both discriminator characteristics at the origin is approximately equal to $D(0) \cong -2$.

Chapter 8

GPS Receivers

A. J. Van Dierendonck
AJ Systems, Los Altos, CA 94024

I. Generic Receiver Description

A. Generic Receiver System Level Functions

A SYSTEM level functional block diagram of a generic GPS receiver is shown in Fig. 1. The generic receiver consists of the following functions: 1) antenna; 2) preamplifier; 3) reference oscillator; 4) frequency synthesizer; 5) downconverter; 6) an intermediate frequency (IF) section; 7) signal processing; and; 8) applications processing.

In a general sense, not all GPS receivers perform navigation processing. Many perform time transfer or differential surveying, or simply collect measurement data. Thus, the last function is more appropriately called applications processing, thus covering a broad set of applications.

The antenna may consist of one or more elements and associated control electronics, and may be passive or active, depending upon its performance requirements. Its function is to receive the GPS satellite signals while rejecting multipath and, if so designed, interference signals. The preamplifier generally consists of burnout protection, filtering, and a low-noise amplifier (LNA). Its primary function is to set the receiver's noise figure and to reject out-of-band interference.

The reference oscillator provides the time and frequency reference for the receiver. Because GPS receiver measurements are based on the time-of-arrival of pseudorandom noise (PRN) code phase and received carrier phase and frequency information, the reference oscillator is a key function of the receiver. The reference oscillator output is used in the frequency synthesizer, from which it derives local oscillators (LOs) and clocks used by the receiver. One or more of these LOs are used by the downconverter to convert the radio frequency (rf) inputs to intermediate frequencies (IFs) that are easier to process in the IF section of the receiver.

The purpose of the IF section is to provide further filtering of out-of-band noise and interference and to increase the amplitude of the signal-plus-noise to a workable signal-processing level. The IF section may also contain automatic

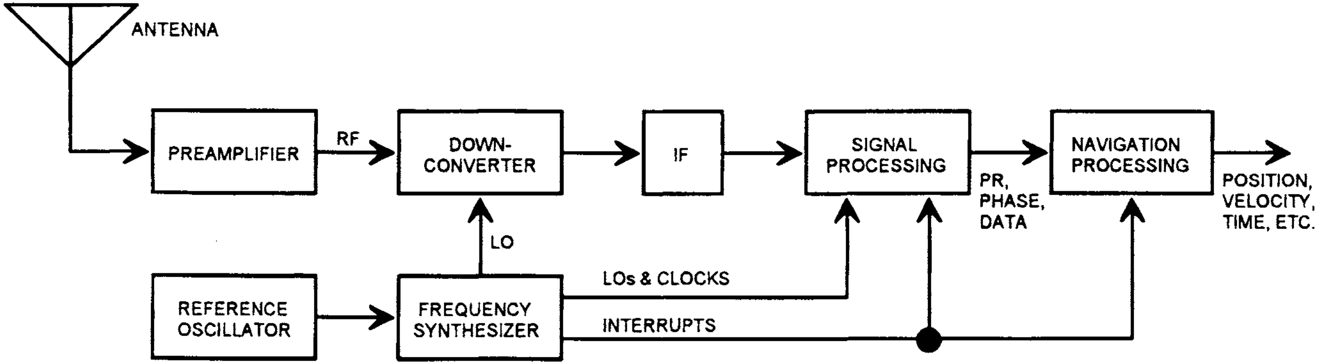


Fig. 1 Generic GPS receiver functional block diagram.

gain control (AGC) circuits to control that workable level, to provide adequate dynamic range, and to suppress pulse-type interference.

The signal-processing function of the receiver is the core of a GPS receiver, performing the following functions:

1) splitting the signal-plus-noise into multiple signal-processing channels for signal-processing of multiple satellites simultaneously;

2) generating the reference PRN codes of the signals;

3) acquiring the satellite signals;

4) tracking the code and the carrier of the satellite signals;

5) demodulating the system data from the satellite signals;

6) extracting code phase (pseudorange) measurements from the PRN code of the satellite signals;

7) extracting carrier frequency (pseudorange rate) and carrier-phase (delta pseudorange) measurements from the carrier of the satellite signals;

8) extracting signal-to-noise ratio (SNR) information from the satellite signals; and

9) estimating a relationship to GPS system time.

The outputs of the signal-processing function are pseudoranges, pseudorange rates and/or delta pseudoranges, signal-to-noise ratios, local receiver time tags, and GPS system data for each of the GPS satellites being tracked, all of which are used by the applications-processing function. Most of this chapter is devoted to the details of the signal-processing function.

The applications-processing function controls the signal-processing function and uses its outputs to satisfy application requirements. These requirements vary with application. Although GPS is primarily a satellite navigation system, the applications of a GPS receiver are diverse. Some of the other applications with significantly differing processing requirements are as follows:

1) time and frequency transfer;

2) static and kinematic surveying;

3) ionospheric total electron content (TEC) and amplitude and phase scintillation monitoring;

4) differential GPS (DGPS) reference station receivers; and

5) GPS satellite signal integrity monitoring.

The common link between these diverse applications is that they all use the same signal-processing measurements in one form or another. However, because of bandwidth and accuracy requirements imposed by these various applications, the requirements on signal-processing function also differ. In general, GPS receivers do not meet the signal-processing and applications-processing requirements for all the applications. Special processing is required for some of the applications.

B. Design Requirements Summary

In addition to the variation in signal- and applications-processing requirements with application, the receiver front-end functions will also have varying design requirements with application. In the following, these variations are noted by function, with details provided later in the chapter.

1. Antenna

The parameters that dictate the antenna requirements are as follows: gain vs azimuth and elevation, multipath rejection, interference rejection, phase stability and repeatability, profile, size, and environmental conditions. The gain requirements are a function of satellite visibility requirements and are closely related to multipath rejection, and somewhat related to interference rejection. The goal is to have near uniform gain toward all satellites above a specified elevation angle, but, at the same time, reject multipath signals and interference typically present at low elevation angles. These are usually conflicting requirements. Some multipath rejection can also be achieved by reducing the left-hand, circularly polarized (LHCP) gain of the antenna without reducing the right-hand, circularly polarized (RHCP) gain. This is because the satellite signals are RHCP signals; whereas, reflected multipath signals usually tend to be either linearly polarized (LP) or even LHCP, depending upon the dielectric constant of the reflecting surface.

Interference rejection can also be achieved using a phased-array antenna, where the relative phase received from each antenna is controlled to “null” out the interference in the combined reception. This type of antenna is called a controlled-reception pattern antenna (CRPA), which is usually used only for military applications.

Phase stability and repeatability are important in differential surveying applications when differential carrier-phase accuracy is important. In this case, orientation of the antenna is important, taking advantage of phase repeatability.

Antenna profile is important in dynamic applications, such as for aircraft and missiles. Normally, requiring a low profile for those applications must be traded off against a desired gain pattern or other desired parameter.

Environmental conditions dictate the type of material used for the antenna and whether or not a radome is required. Some materials change their dielectric properties as a function of temperature.

2. Preamplifier

The preamplifier generally consists of burnout protection, filtering, and an LNA. The parameters that dictate the preamplifier requirements are as follows: the unwanted rf environment as received through the antenna, losses that precede and follow the preamplifier, and desired system noise figure (or noise temperature) as derived from overall receiver performance requirements. The gain of the preamplifier is not a system-level requirement, per se, but a derived requirement that satisfies that system level requirement.

The unwanted rf environment as received through the antenna affects the preamplifier in two ways. Either it could cause damage to the preamplifier electronics, or cause saturation of the preamplifier and circuitry that follows. Of course, except for damage prevention, we can do nothing to suppress the rf environment, as passed by the antenna, at frequencies that are in the bandwidth of the desired GPS signal. (Actually, there are techniques such as adaptive equalizers, usually applied at IF. These techniques are not discussed in this chapter.) That environment is considered to be either jamming or unintentional interference. However, suppression of the rf environment out of the desired GPS signal band can be accomplished by filtering, either before, during, and/or after amplification. When it is accomplished, it is based upon a trade-off between system noise figure requirements and filter insertion loss and bandwidth efficiency. Suppression of in-band and out-of-band damaging interference is usually accomplished with diodes that provide a ground path for strong signals. In the case of lightning protection, more complex lightning arrestors are sometimes used.

The system noise figure is set using an LNA that provides enough gain to cause any losses inserted after the LNA to have a negligible effect. Losses inserted prior to the LNA add directly to the system noise figure and are not affected by the LNA.

3. *Reference Oscillator*

The requirements on reference oscillators for GPS receivers have changed considerably over the years, mainly because of their expense. A high-quality oscillator can be the most significant cost item of a modern receiver. Thus, there have been compromises made on oscillator performance. Also, the oscillator's performance is not as critical in the modern multichannel receivers, especially in most commercial applications. However, there are some commercial and military applications where reference oscillator performance is critical. Typical requirements applied to reference oscillators are as follows:

1. *Size*—Stable oven-controlled crystal oscillators (OCXOs) and rubidium oscillators can be relatively large. Temperature-compensated crystal oscillators (TCXOs) are relatively small. Larger oscillators have more temperature inertia.
2. *Power*—Oven-controlled crystal and rubidium oscillators consume significant power.
3. *Short-term stability* caused by temperature, power supply, and natural characteristics. Short-term stability affects the ability to estimate and predict time and frequency in the receiver.
4. *Long-term stability* caused by natural characteristics, including crystal aging
5. *Sensitivity to acceleration*—*g* force and vibration sensitivity. Vibration causes phase noise, and dynamic *g* forces affect the ability to estimate time and frequency in the receiver.
6. *Phase noise*—high-frequency stability. Phase noise degrades the signal-processing performance of the receiver.

4. *Frequency Synthesizer*

Mostly, the requirements placed on the frequency synthesizer are derived requirements and the receiver designer's choice. Its design is based on the designer's *frequency plan,* which defines the receiver's IF frequencies, sampling clocks, signal processing clocks, etc. The frequency plan requires careful analysis to ensure adequate rejection of mixer harmonics, LO feed-through, unwanted sidebands and images. A key design parameter for the synthesizer is the minimization of phase noise generated in the synthesizer. Phase noise generated at the reference oscillator frequency is multiplied by the ratio of the rf frequency to its frequency through the synthesis process. Thus, the design of the synthesizer is critical to the performance of the GPS receiver.

The frequency synthesizer may also be required to generate local clocks for signal processing and interrupts for applications processing. This requirement might be assigned to signal processing. These local clocks comprise the receiver's time base.

5. *Downconverter*

The downconverter mixes LOs generated by the frequency synthesizer with the amplified rf input to IF frequencies, and, if so designed, IF frequencies to lower IF frequencies. This process implements the frequency plan, which, again, is the receiver designer's choice. The outputs of the mixers include both the lower and upper sideband of the mixing process, either of which can be used as an IF frequency. The unwanted sideband, LO feed-through, and harmonics are rejected by filtering at the IF. The unwanted image is filtered at rf before the mixing process. Because all of these processes are a function of the frequency plan, the requirements placed upon the downconverter are also derived requirements.

6. *Intermediate Frequency Section*

The requirements on the IF section are as follows:

1) Final rejection of out-of-band interference, unwanted sidebands, LO feed-through, and harmonics. The bandwidth of this rejection is a trade-off against correlation loss caused by filtering. In addition, the rejection of wide-band noise is required to minimize aliasing in a sampling receiver.

2) Increase the amplitude of the signal-plus-noise to workable levels for signal processing and control that amplitude as required for signal processing (AGC).

3) Suppress pulse-type interference.

4) Depending upon design, convert the IF signal to a baseband signal composed of in-phase (I) and quadraphase (Q) signals.

7. *Signal Processing*

As previously stated, signal processing may include the generation of the local clocks and interrupts. However, its prime requirement is to provide the GPS measurements and system data from selected satellites required to perform the navigation or other applications function. How this requirement is met constitutes the signal–processing-derived requirements and is based upon the receiver designer's choice. The functions of the signal processing are listed in the beginning of this section.

8. *Applications Processing*

The requirements for applications processing are to control the signal processing to provide the necessary measurements and system data and to use those measurements and system data to perform one or more of a variety of GPS applications. The processing requirements for these applications are covered elsewhere in this volume.

II. Technology Evolution

A. Historical Evolution of Design Implementation

In this section, the historical evolution of GPS receiver design implementation is described, dating back to the mid-1970s, the start of GPS Phase I concept validation. This does not include the development effort during the 621B program, predecessor to the GPS.

The first seven GPS receivers were mostly developed concurrently. These were the Phase I sets—the X Set, the Y Set, the Z Set, the Manpack developed by Magnavox, the High Dynamic User Equipment (HDUE) developed by Texas Instruments, the Advanced Development Model (ADM) developed by Collins Radio, and a satellite-monitoring receiver developed by Stanford Telecommunications, Inc. Three of these receivers, the Z Set, Manpack, and the satellite-monitoring receiver, had analog baseband signal processing and used processors only for the applications-processing function. The other three used microprocessors of one form or another to perform some of the signal processing. Needless to say, these receivers were all quite large, except for the medium sized Z Set and the Manpack. However, these two receivers had only one channel with one correlator, using this channel in a sequencing mode.

The rf sections of these receivers used discrete components, such as transistors, including large cavity filters. Circuit isolation was difficult. Consequently, there were a number of IF stages (three or four or more) to distribute the gain over a number of frequencies. This made the frequency synthesizers very complicated. Furthermore, the high-frequency LOs were generated using cavity multipliers.

The lower-frequency analog signal-processing sections were made up of operational amplifiers and other discrete components. The digital portions, such as the code generators and clocks, consisted of medium-scale integrated (MSI) circuits, at best. CMOS circuitry was in its infancy at that time, so higher-power consumption circuitry was used. The X Set used a bit-slice process controller for its baseband signal processing, which was state of the art at that time. Computers ranged from Hewlett Packard minicomputers to DEC large-scale integrated (LSI) computers. A structured version of Fortran was used by Magnavox as the applications-processing language; whereas, assembler language was used for the baseband signal processing.

The evolution into the Phase II GPS receivers was not that dramatic, although analog baseband signal processing was completely replaced with microprocessors and some MSI gave way to LSI. The evolution into the Phase III GPS receivers included some digital gate arrays and more powerful microprocessors. However, multiple IF stages and complicated frequency plans were still common. Signal processing was still not accomplished digitally, and the receivers were generally still quite large and expensive. Some of these receivers are still being produced today (the RCVR-3A airborne set and the RCVR-3S shipboard set). The operational control segment (OCS) monitor stations still use large hybrid analog/digital baseband receivers. However, these receivers are no longer in production.

B. Current Day Design Implementation

Except for the RCVR-3A and the RCVR-3S, all GPS receivers in production today are *probably* all true digital signal-processing receivers. I use the word *probably* because there are so many different GPS receivers in existence today, that one could never know. However, the consensus is that, in order to produce inexpensive receivers, they have to perform true digital signal processing. One reason for this is because almost all new GPS receivers have at least 4 full tracking channels; whereas, some have up to 24 to 36 channels for dual-frequency processing. Even the new military receivers, which usually lag in technology, have up to six tracking channels. The evolution of CMOS very-large scale integration (VLSI) technology has caused an explosion in the capabilities of new GPS receivers. The two technologies were made for each other. The processing speed of the new CMOS matches the signal-processing requirements for GPS receivers. Five or more channels on a chip are the norm in new GPS receivers.

The front-end electronics of GPS receivers have also experienced a dramatic evolution with the introduction of monolithic microwave integrated circuits (MMIC), stripline filter techniques, chip capacitors and resistors, high-speed digital integrated circuits, surface acoustic wave (SAW) filters, and surface mount printed circuit boards (PCBs).

1. Radio Frequency Electronics

The use of gallium arsenide (GaAs) MMIC for GPS receivers has been attempted and used on occasion, but with limited success. Most modern receivers use silicon bipolar technology because of efficiency and cost. Although the Defense Advanced Research Projects Agency (DARPA) sponsored Rockwell Collins on a program for the development of gallium arsenide (GaAs) MMIC for GPS receivers, Rockwell abandoned its use in favor of silicon bipolar in their miniature airborne GPS receiver (MAGR) design for those very reasons.[11] The GPS frequencies are not high enough to warrant the use of highly integrated GaAs circuitry. Most rf designers agree that the use of GaAs is usually limited to a front-end FET, and little more than a marketing "buzz word." Any advantage gained in using GaAs with respect to noise figure and power dissipation is lost because of circuit-matching problems at the GPS frequencies.

To an extent, the same is true for large-scale MMIC, even if silicon bipolar is used. However, in this case, the use of large-scale MMIC is a trade-off between size and circuit efficiency. Large-scale MMIC is certainly smaller, but may not be as efficient as smaller-scale MMIC in terms of power dissipation, cost, and out-of-band rejection. That is because the large-scale MMIC is designed for a wide range of applications, not just for a GPS receiver. Furthermore, in the rf and IF sections of a receiver, filters must be injected between stages of amplification and mixing for out-of-band rejection and gain stability. Efficient filtering cannot be accomplished in the MMIC chips.

Times are changing, however. GEC Plessey has introduced a silicon bipolar MMIC chip, the GP1010, that was developed specifically for GPS receivers.[2] Its input is at rf at the GPS L_1 frequency, although already amplified and filtered via a preamplifier. Intermediate frequency filtering is not included on the chip. Its output is a stream of 1.5-bit samples for input to digital signal processing. The chip also contains the frequency synthesizer and AGC, but not the loop filters.

2. *Frequency Synthesizer Electronics*

The use of cavity multipliers has been replaced with phase–lock-loops (PLLs) using voltage-controlled oscillators (VCOs) and high-speed digital divider circuits (prescalers). Chips are commercially available that contain programmable prescalers. However, because of the clocking speeds required for the first LO provided by these frequency synthesizers, sometimes the frequency of the VCO and clocking speed of the prescaler are half what is required, and a frequency doubler or a second-harmonic mixer is used to achieve the *L*-band LO. However, these components are still small and quite simple. In the end, we have small, efficient frequency synthesizer electronics that are significantly smaller and lower power than technology used 5–10 years ago. As described above, the GEC Plessey GP1010 chip contains the VCO and prescalers, but it is not programmable.[2]

3. *Down Conversion and Intermediate Frequency Electronics*

The conversion from rf to IF has generally become a silicon bipolar MMIC implementation, where the MMIC includes mixing and a stage or two of amplification. The VCO for the LO may also be included. Once at IF, surface acoustic wave (SAW) filters provide a small, efficient means of final filtering to minimize unwanted out-of-band signals and noise.

4. *Reference Oscillators*

Unfortunately, the development of good, stable reference oscillators has not kept up with the pace of the development of the other sections of a GPS receiver. There have been improvements, but, depending upon the ultimate stability required, the oscillator can be the most expensive and largest component in the receiver. For example, for very good stability, not including that achieved with relatively large and expensive atomic oscillators, a crystal oscillator must be ovenized to minimize frequency excursions that are the main cause of frequency instability. Unfortunately, to have temperature inertia, mass is required. Thus, although the sizes of these oscillators have been reduced over the years, they are still relatively large. In fact, these smaller oscillators are not usually as stable as the older, larger, oscillators.

The advancement of digital signal processing and multiple receiver channels has allowed the use of TCXOs in most commercial applications. This has prompted some improvement in TCXO performance over what had been the norm in the past. However, that improvement has not been dramatic, although the size and cost of the TCXOs have been reduced significantly.

III. System Design Details

In this section, the block diagram of Fig. 1 is expanded, and design details are presented, starting with the hardware and following with the software part of the digital signal processing. Only the implementation of modern digital receivers is presented along with some trade-offs between different modern implementations. Specifically, two receiver designs that are familiar to the author are used as examples—the Rockwell Collins miniature airborne GPS receiver

(MAGR), an example of a military receiver, and the NovAtel GPSCard™, an example of a commercial receiver. A functional overview of the MAGR is shown in Fig. 2.[1] Details of these functions may vary from receiver to receiver, but they exist in one form or another in all modern receivers. Although the MAGR is a dual-frequency receiver that receives both L_1 and L_2 signals and both the C/A- and P-codes, processing that is common for those signals is presented only once.

A. Signal and Noise Representation

Before describing the remainder of the receiver operations, it is appropriate to provide a representation of the received signal and noise in both the time domain and the frequency domain. The signal is represented as follows:

$$s(t) = AC(t)D(t)\cos[(\omega_0 + \Delta\omega)t + \phi_0] \tag{1}$$

where A = signal amplitude; $C(t)$ = PRN code modulation (± 1); $D(t)$ = 50 bps data modulation (± 1); $\omega_0 = 2\pi f_0$ = carrier frequency (L_1 or L_2); $\Delta\omega = 2\pi\Delta f$ = frequency offset (Doppler, etc.); and ϕ_0 = nominal (but ambiguous) carrier phase.

For the purpose of the processing described herein, it is necessary to represent only one component of the L_1 signal, because one of the two components (in-phase or quadrature) does not correlate with the receiver channel's reference code in channel tracking the desired code. L_2 processing is identical to that of the L_1 signal except that L_2 acquisition and tracking are usually aided with information obtained by tracking the more powerful L_1 signal.

In the frequency domain, the spectral density of the signal is the spectral density of the PRN code centered at $\pm(\omega + \Delta\omega)$. At baseband, this spectral density is as follows:

$$S_s(\omega) = \frac{A^2 T_c}{2} \frac{\sin^2(\omega T_c/2)}{(\omega T_c 2)^2} \tag{2}$$

where T_c is the PRN code chip width, or the inverse of the PRN code chipping rate. Although the C/A-code spectrum is a line spectrum, this representation suffices for most of the processing described herein. Exceptions are noted as they arise.

Signal power in a $2B$ Hz two-sided bandwidth is given by the following:

$$P_s = \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} S_s(\omega) \mathrm{d}\omega \tag{3}$$

If $B = \infty$, $P_s = A^2/2$. If $B = 1/T_c$, $P_s = 0.9\, A^2/2$, resulting in a 0.45 dB signal loss.

Ambient noise is represented in the frequency domain as white noise with a constant spectral density $N_0/2$. If passed through a unity-gain bandpass filter with a two-sided noise bandwidth $2B$ Hz centered at frequency ω_0, the spectral density appears as illustrated in Fig. 3. The resulting noise power in the two-sided bandwidth is as follows:

$$P_n = 2N_0 B \tag{4}$$

and a signal-to-noise ratio in a $2B$ two-sided noise bandwidth of

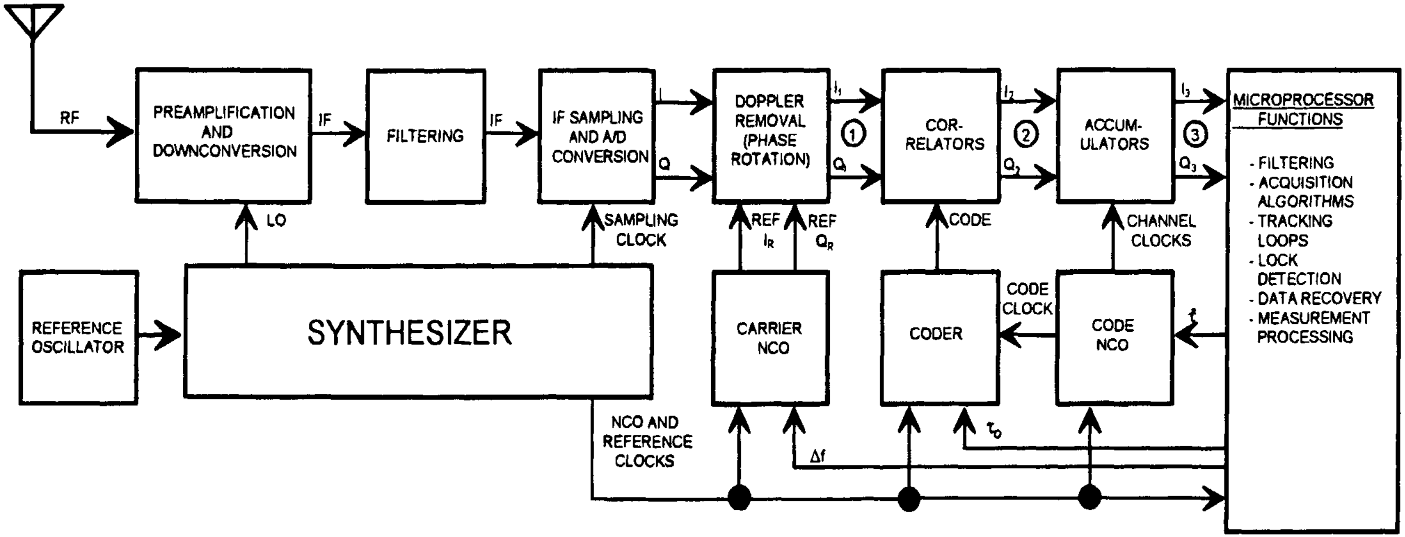


Fig. 2 Functional block diagram of the MAGR.

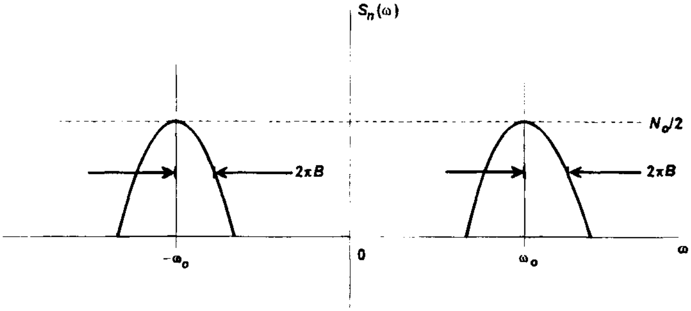


Fig. 3 Spectral representation of ambient noise.

$$\frac{P_s}{P_n} = \frac{A^2}{4N_0B} \tag{5}$$

The output of the bandpass filter has two equivalent time domain representations[3]:

$$n(t) = x(t)\cos\omega_0 t - y(t)\sin\omega_0 t \tag{6}$$

$$n(t) = r(t)\cos[\omega_0 t + \varphi(t)] \tag{7}$$

where $x(t)$ and $y(t)$, the in-phase and quadraphase components, respectively, are bandlimited Gaussian processes with the properties defined as follows:

$$\begin{aligned} E[x(t)] &= E[y(t)] = 0 \\ E[x^2(t)] &= E[y^2(t)] = E[n^2(t)] = P_n \\ E[x(t)y(t)] &= 0 \end{aligned} \tag{8}$$

Furthermore, $r^2(t) = x^2(t) + y^2(t)$ has a chi-squared (or Rayleigh) distribution with two DOF, and

$$\varphi(t) = \tan^{-1}\left[\frac{x(t)}{y(t)}\right] \tag{9}$$

is uniformly distributed between 0 and 2π radians.

B. Front-End Hardware

The hardware is comprised of the front-end electronics and part of the digital signal processing. The other part of the digital signal processing is implemented in software.

1. Antenna

The following requirements were placed upon the Fixed Reception Pattern Antenna (FRPA3) for Rockwell's military receivers[4]:

1) *rf:* "The GPS antenna shall accept the GPS navigation signals at both the L_1 and L_2 frequencies and output them to the GPS antenna electronics."[4] This implies either a wide-band antenna that accepts both frequencies and all frequencies between them, or implies two antennas packaged in a single unit. Both types of antennae are available, because some are naturally narrow band while others are wide band.

2) *Antenna gain:* "The GPS antenna shall provide a minimum gain of −2.5 dBic to a RHCP signal over a 160° solid angle cone of coverage (above 10° elevation angle) for signals in both the L1 and L2 bandwidths. Gain shall be measured at the L1 and L2 carrier frequencies at the prevailing ambient temperature using a standard gain horn for comparison. The combined effects of environmental temperature range and bandwidth . . . shall not cause the gain to be less than −3.3 dBic. These gain requirements apply when using a test ground plane."[4] (This ground plane matches the housing of the antenna electronics for shipboard installation.)

3) *Other antenna specifications:* Other FRPA3 specifications are as follows:

a) *Voltage standing wave ratio (VSWR):* ≤2:1 (referenced to 50 ohms over L_1 and L_2 bands)

b) *dc impedance @ signal interface:* 0 ohms

c) *Connector:* Single TNC female receptacle

d) *Size:* Less than 6 in. in diameter and 1.75 in. in height, including connector

The dc impedance of 0 ohms provides lightning protection. More than one vendor supplies the antenna and not necessarily through Rockwell Collins. Typical gain patterns from one vendor (Sensor Systems, Inc. S67-1575-14) are shown in Fig. 4. This model has two narrowband antennas packaged in a single unit. The Sensor Systems antenna has a diameter of 3.5 in. and a height of 0.565 in.[5]

The following technical specifications are given for a Sensor Systems' commercial aeronautical antenna (S67-1575-16), which was designed to meet ARINC 743A characteristics[5,6]:

1) *rf:* 1575 ±2 Mhz with a VSWR ≤ 1.5:1 or 1575 ±10 MHz with a VSWR ≤2:1.

2) *antenna gain:* The gain pattern is specified as follows:

> −1 dBic to 75 deg from vertical
> −2.5 dBic to 80 deg from vertical
> −4.5 dBic to 85 deg from vertical
> −7.5 dBic at 90 deg from vertical

The RHCP gain pattern is shown in Fig. 5.

2. *Receiver Front End*

The receiver front end consists of filtering and limiting, an LNA, a frequency synthesizer, downconversion, and conversion to baseband. Initial filtering, limiting, and LNA can be housed with the antenna to comprise an integrated antenna electronics. This is an optional configuration of the MAGR and is usually the case in commercial receivers. In the case of the MAGR, the following requirements are imposed:[7]

1. *Preselector filtering:* The preselector filtering is required to reject out-of-band interference and to limit the noise bandwidth of the antenna electronics.

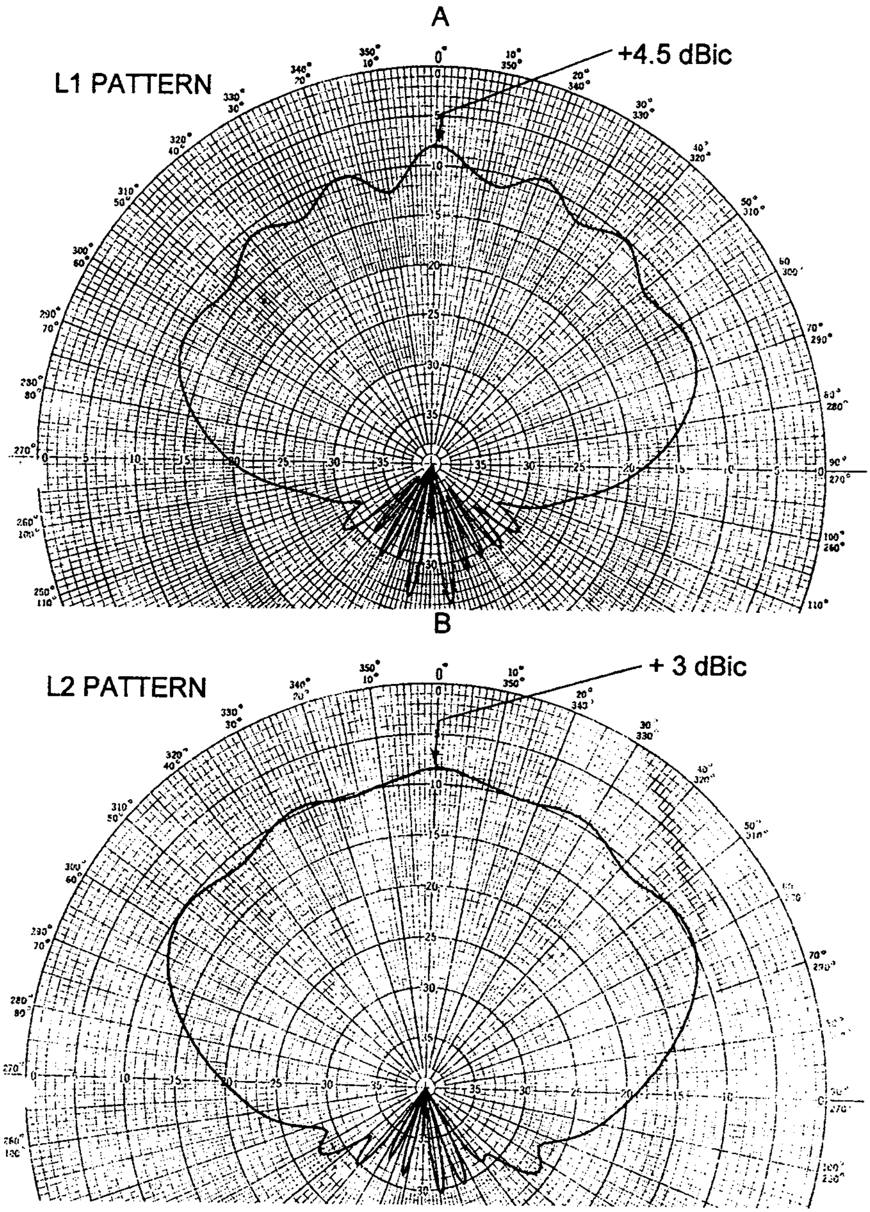


Fig. 4 Fixed Reception Pattern Antenna 3 antenna gain patterns (courtesy of Sensor Systems, Inc.).

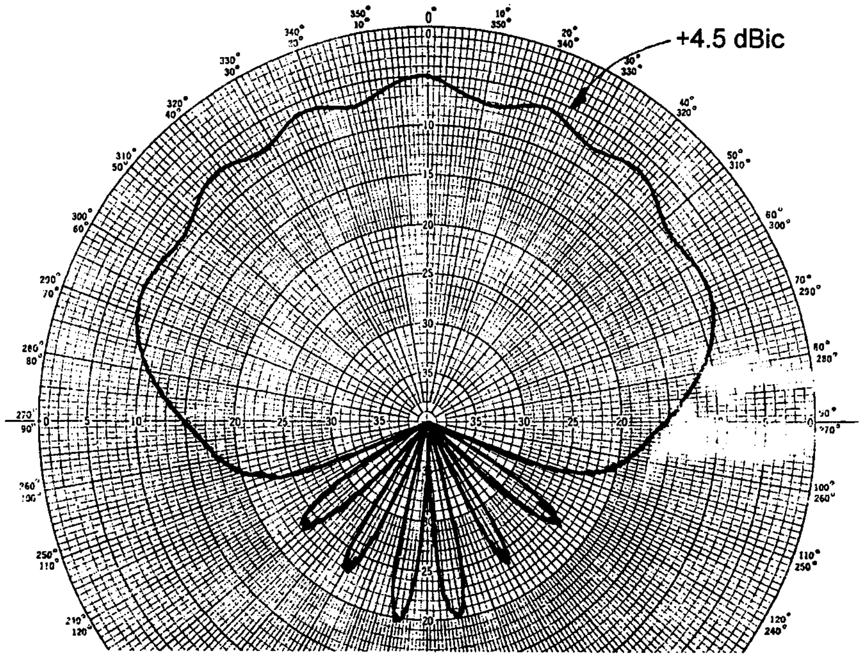


Fig. 5 Commercial L_1 Antenna Gain Patterns (courtesy of Sensor Systems, Inc.).

The filtering shall be dual-band centered at L_1 and L_2, with a noise bandwidth of 80 MHz in each band. The insertion loss shall be sufficiently low to meet overall gain and noise figure requirements.

2. *Burnout protection:* The antenna electronics shall not incur damage or performance degradation after being subject to a peak signal power density of 69 kW/m^2 for not more than 10 μs, or a continuous signal power density of 348 W/m^2 in either band of frequencies.

3) *Gain and noise figure:* The antenna electronics, including interconnecting cabling, preselector, and protection circuitry, shall have a minimum overall gain of 23 dB, with a noise figure of 4 dB at the input to the MAGR receiver.[8] The maximum overall gain shall be 33 dB.

The Sensor Systems' commercial ARINC 743A antenna described is also available with an internal preamplifier with 26 ±3 dB gain with an internal interference rejection filter and an LNA (S67-1575-52). A single TNC connector carries both the L_1 signal and dc power to the LNA (+4 to +24 VDC at 25 mA, maximum). Out-of-band rejection is 35 dB at 1625 MHz.

3. *Noise Figure Computations*

Figure 6 provides a model of a receiver front end for the purposes of computing the receiver and system noise figure and noise temperature. In general, the system noise figure (in dB) is related to system noise temperature (in Kelvin) as follows[9]:

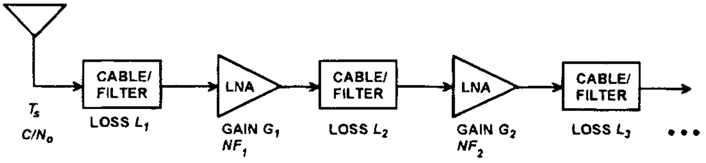


Fig. 6 Noise figure computation model.

$$\mathrm{NF} = 10\log_{10}\left(1 + \frac{T_{\mathrm{sys}}}{T_0}\right) \tag{10}$$

where $T_0 = 290$ K $= 24.6$ dB−K.

The corresponding noise density, in W/Hz, is $N_0 = K_B T_{\mathrm{sys}}$; where $K_B = -228.6$ dBW/K−Hz $= 1.380 \times 10^{-23}$W/K−Hz, is the Boltzmann constant.[9] Based upon the model given in Fig. 6, the system noise temperature is computed as follows:[10,11]

$$\begin{aligned} T_{\mathrm{sys}} &= T_s + T_R \\ &= T_s + T_0[L_1 - 1 + L_1[\mathrm{NF}_1 - 1 + G_1^{-1}[L_2 - 1 \\ &\quad + L_2\,[\mathrm{NF}_2 - 1 + G_2^{-1}[L_3 - 1 + \cdots]]]]] \end{aligned} \tag{11}$$

where T_s is the source (antenna) temperature; T_R is the receiver noise temperature; and the L_i, NF_i, and G_i are loss, amplifier noise figure, and amplifier gain of each stage i, respectively, all given in ratio. This formula is known as the *Friis Formula.* Note that the loss and noise figure of the first stage affects the system noise figure directly; whereas, any losses after the first amplification are reduced proportional to that gain. Thus, the first stage is said to *set* the system noise figure or noise temperature.

Note, also, that the source temperature adds directly to the system noise temperature. Normally, this is the antenna sky temperature, which is relatively low with respect to the ambient noise temperature T_0. One exception is when the receiver is connected directly to a GPS signal generator or simulator, in which the source temperature is the ambient noise temperature (290 K), which is a worst-case situation. Normally, adjustments to the signal power must be made to compensate for this. Note that in this case, if the first stage gain is high enough, the noise density is simply as follows:

$$\begin{aligned} N_0 &= K_B T_0 \mathrm{NF}_1 \\ &= -228.6 + 24.6 + \mathrm{NF}_1(\mathrm{dB}) \\ &= -204\ \mathrm{dBW/Hz} + \mathrm{NF}_1(\mathrm{dB}) \end{aligned} \tag{12}$$

Sometimes this equation is erroneously used for "real-world" computations, providing pessimistic analysis results. A source temperature of 75–100 K is typical, depending upon the antenna pattern and the amount of ground temperature observed.[9,12]

4. *Synthesizers and Frequency Plans*

Figure 7 presents the block diagram of two of the MAGR's custom silicon bipolar chips that make up its synthesizer.[1] One chip (*L*-band chip) also includes

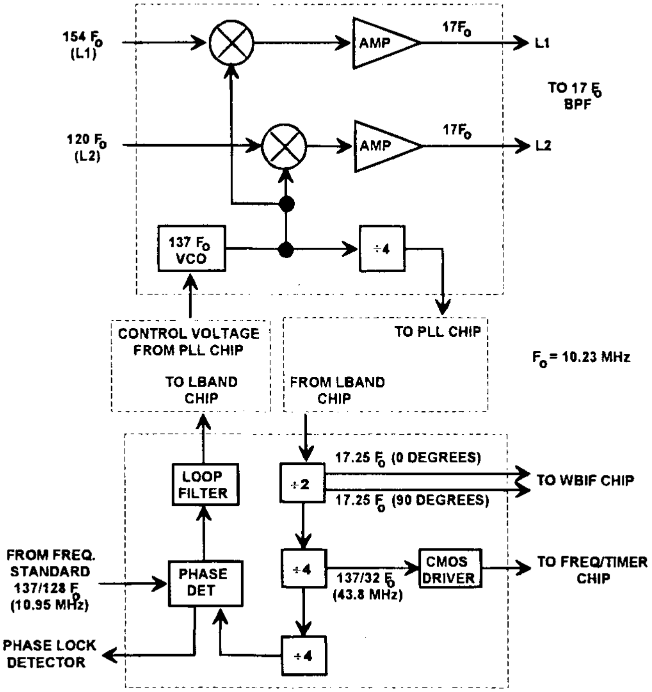


Fig. 7 Miniature airborne GPS receiver *L*-band/phase-lock-loop chips and frequency plan.

the downconverter function of the receiver. This synthesizer design reflects the concept used in modern GPS receivers, where a nonstandard reference oscillator frequency is used. Older receivers used such frequencies as 5 or 10 MHz, or 5.115 or 10.23 MHz, which also became standards because of GPS. Note that the MAGR uses 10.949296875 MHz as its reference.

The MAGR synthesizer generates a common local oscillator at $137F_0$ (1401.51 MHz) for both the L_1 and L_2 frequencies by phase-locking the LO voltage controlled oscillator, divided by 128, to the 10.95 MHz reference (F_0 = 10.23 MHz). It also generates common in-phase and quadraphase LOs for conversion of the IF frequencies to baseband at 17.25 F_0 (176.4675 MHz) and a CMOS clock at 43.7971875 MHz, which is used for clocking the digital signal-processing circuitry. Note that the baseband LOs do not match the IF frequecies at 17 F_0. This leaves a residual frequency offset at baseband, which is part of the frequency planning scheme. The overall effects of this residual offset is treated as Doppler. These effects are described later. However, along with the nonstandard reference frequency, allowing for such an offset also simplifies the frequency plan and the synthesizer design.

The GPSCard™ uses the same concept for its frequency plan and synthesizer as illustrated in Fig. 8. Its reference frequency is 20.473 MHz, which is also its digital signal-processing clock. The GPSCard™ utilizes a commercial synthesizer

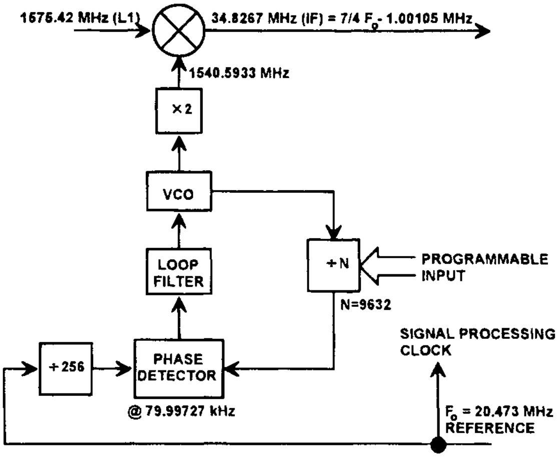


Fig. 8 GPSCard™ synthesizer and frequency plan.

chip with a programmable divider (prescaler) rather than custom chips. The key difference between this frequency plan and that of the MAGR is that the GPSCard™ uses IF sampling for its conversion to baseband, which is described later. Note that there is also a residual frequency offset.

5. *Mixing Operations and Intermediate Frequency Filtering*

Downconversion from rf to IF, and, in the case of the MAGR, conversion from IF to baseband, are accomplished by mixing the incoming signal and noise (rf or IF) with an LO. This process is illustrated in Fig. 9.

If the local oscillator is represented as $LO_1(t) = 2\cos\omega_1 t$ with power of two units, then the output of the mixer is this LO multiplied by the sum of Eq. (1) and Eq. (7), or $s_{\mathrm{IF}}(t) + n_{\mathrm{IF}}(t) = 2[s(t) + n(t)]\cos\omega_1 t$ + harmonics + LO feedthrough + image noise. Ignoring for the moment the harmonics, LO feedthrough, and image noise and using the product of cosines, for the signal we have $s_{\mathrm{IF}}(t) = AC(t)D(t)\{\cos[(\omega_0 + \omega_1 + \Delta\omega)t + \phi_0] + \cos[(\omega_0 - \omega_1 + \Delta\omega)t + \phi_0]\}$ consisting of upper and lower sideband components, each with a power of

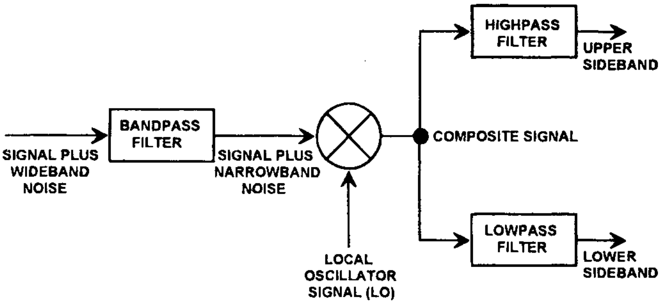


Fig. 9 Mixing operations.

$A^2/2$. Only the lower sideband is wanted. Therefore, the upper sideband is eliminated via a low-pass filter, resulting in the IF frequency $\omega_{\mathrm{IF}} = \omega_0 - \omega_l$. Similarly, for the noise, in terms of Eq. 7, $n_{\mathrm{IF}}(t) = r(t)\cos[\omega_{\mathrm{IF}}t + \varphi(t)]$.

Harmonics and LO feedthrough are removed via a well-designed frequency plan and the use of bandpass filters at IF. Harmonics are generated by the mixer, because it is a nonlinear device. LO feedthrough (at ω_l) is the LO leaking through the mixer. In addition, these filters also provide the final rejection of out-of-band interference and image noise, primarily because it is easier to obtain narrowband filtering at the lower IF frequencies.

In-band image noise is either at the frequency $\omega_0 - 2\omega_l$ (upper sideband component at the lower sideband), or at the frequency $\omega_0 - 2\omega_{\mathrm{IF}}$ (mixes to the negative of the lower sideband). Both would mix to the IF frequency and, thus, would not be filtered at IF. It is necessary to filter noise at these frequencies prior to mixing at rf so that they do not exist. In order to avoid the use of narrowband filters at rf for this purpose, the frequency plan should be designed to prevent image noise close to the IF frequency.

6. *Conversion to Baseband*

Conversion to baseband is the process of converting the IF signal to that of in-phase and quadraphase components of the signal envelope, but still modulated with residual Doppler. However, as pointed out in the discussions of synthesizers and frequency plans, in most modern receivers, an intentional residual frequency offset may still exist. There are two methods for achieving this conversion—by analog mixing or by a technique known as IF (or pass-band) sampling. Because the latter (used in the GPSCard™) is a sampling process, its description is delayed to the next section.

The MAGR uses the former method, and does so in its wide-band IF chip (silicon bipolar), which is shown in Fig. 10.[1] Also shown are the AGC and the analog-to-digital (A/D) converters. This conversion to baseband is realized by mixing the IF signal with two LOs, one of which is shifted 90° in phase with respect to the other (in quadrature). The low-pass filters reject the upper sidebands. The in-phase and quadraphase combined unity power LOs are, respectively,

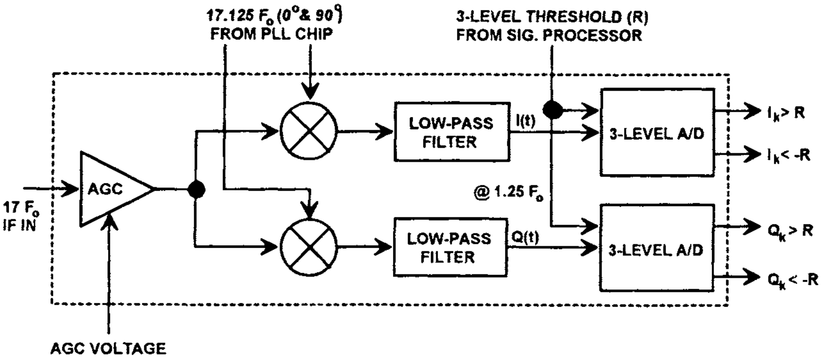


Fig. 10 Miniature airborne GPS receiver wideband intermediate frequency chip.

as follows:

$$\mathrm{LO}_{2I}(t) = \sqrt{2}\cos\omega_2 t \tag{13}$$

$$\mathrm{LO}_{2Q}(t) = \sqrt{2}\cos\left(\omega_2 t + \frac{\pi}{2}\right) = -\sqrt{2}\sin\omega_2 t \tag{14}$$

The resulting analog in-phase and quadraphase baseband signal components are then as follows:

$$I_s(t) = \frac{A}{\sqrt{2}} C(t)D(t)\cos\left(\Delta\omega_B t + \phi_0\right) \tag{15}$$

$$Q_s(t) = \frac{A}{\sqrt{2}} C(t)D(t)\sin\left(\Delta\omega_B t + \phi_0\right) \tag{16}$$

where the residual frequency offset is

$$\Delta\omega_B = \omega_{\mathrm{IF}} - \omega_2 + \Delta\omega \tag{17}$$

The relationship of these I and Q levels at this point and the signal power, related to that defined in Eq. (3), is $P_s = E[I_s^2(t) + Q_s^2(t)] = A^2/2$.

Under the assumption that this residual frequency offset is quite small with respect to the bandwidth of the low-pass filters, and that their single-sided bandwidths are essentially B Hz, the baseband noise components are simply $I_n(t) = x(t)/\sqrt{2}$ and $Q_n(t) = y(t)/\sqrt{2}$, and the noise power, in terms of Eqs. (4) and (8), is $P_n = E[I_n^2(t) + Q_n^2(t)] = 2N_oB$. The MAGR AGC shown in Fig. 10 does not operate on signal power, because the signal is still below the noise level at this point in the receiver. It is a very wideband AGC whose time constant is such that it suppresses pulse interference (time constant < 1 μs).

C. Digital Signal Processing

Digital signal processing consists of precorrelation sampling, Doppler removal, PRN coders, correlators, number-controlled oscillators (NCOs), postcorrelation filtering, and various receiver clocks.

1. Precorrelation Sampling

As is evident for the MAGR in Fig. 10, modern GPS receivers all become digital prior to correlation and Doppler removal. However, this is where the commonality ends. They differ in sample rates, sample quantization, and, as discussed previously, some receivers convert to baseband as part of the sampling process, known as IF sampling.

Intermediate frequency sampling is illustrated in Fig. 11. The concept is to sample the IF signal at a rate at which the I and Q samples are obtained directly. Suppose the sample rate is as follows:

$$\mathrm{SR} = \frac{4f_{\mathrm{IF}}}{N} \tag{18}$$

where f_{IF} is the IF frequency being sampled, and N is an odd number. Then, the samples would be taken at

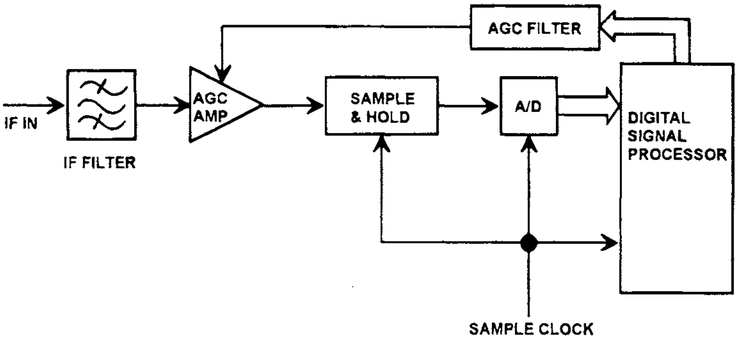


Fig. 11 Intermediate frequency sampling process.

$$t_k = \frac{kN}{4f_{\mathrm{IF}}}\ \mathrm{sec};\ k = 0, 1, \ldots \tag{19}$$

Sampling the IF signal at these times yields the following:

$$\begin{aligned} s_k = s_{\mathrm{IF}}(t_k) &= AC(t_k)D(t_k)\cos\left[2\pi(f_{\mathrm{IF}} + \Delta f)\frac{kN}{4f_{\mathrm{IF}}} + \phi_0\right] \\ &= AC_kD_k\cos\left[\frac{\pi kN}{2}\left(1 + \frac{\Delta f}{f_{\mathrm{IF}}}\right) + \phi_0\right] \\ &= AC_kD_k\cos\left[\frac{\pi kN}{2} + \phi_k\right] \end{aligned} \tag{20}$$

where Δf is an intentional frequency offset plus that attributable to Doppler, C_k and D_k are the code and data at time t_k, and

$$\phi_k = \phi_0 + \frac{\pi kN\Delta f}{2f_{\mathrm{IF}}} = \phi_0 + \Delta\omega\, t_k = \phi_0 + \Delta\phi_k \tag{21}$$

is the baseband phase of the sample attributable to the nominal phase and frequency offset at time t_k. If the sample rate is offset from that of Eq. (18), $\Delta\omega$ in Eq. (21) simply becomes $\Delta\omega_B$, analogous to Eq. (17). If we ignore the Δf of Eq. (20) for the moment, note that the IF signal is sampled at exactly successive 90 deg phases, producing the following sequence of samples:

$$\sqrt{2}[I_{sk}, Q_{sk}, -I_{sk}, -Q_{sk}, I_{sk}, Q_{sk}, -I_{sk}, -Q_{sk}, \ldots] \tag{22}$$

or

$$\sqrt{2}[I_{sk}, -Q_{sk}, -I_{sk}, Q_{sk}, I_{sk}, -Q_{sk}, -I_{sk}, Q_{sk}, \ldots] \tag{23}$$

depending upon the value of N. The Δf results in a time-varying $\Delta\phi_k = \Delta\omega t_k$, causing a phase rotation of the samples that is removed after the sampling process.

This IF sampling process is sometimes called *pseudo sampling,* because the I and Q samples do not occur at the same time. For large frequency offsets with respect to the sampling rate, but still within the Nyquist rate, this will induce an additional phase shift in the Q samples causing aliasing to a negative frequency offset. However, this is not a problem for this GPS receiver application for reasonable Δf values.

The digital signal-processor can simply invert the sign on half the samples and sort them into in-phase and quadraphase samples. This sign inversion is actually an advantage in that, if there are any dc biases present in the sampling process, they will eventually cancel in subsequent signal processing. Also, because the I and Q samples are generated in the same circuitry, there are no gain and phase imbalances between them, except as noted above. This can occur in the analog baseband conversion process. Also, only one A/D converter is required, although it must sample at twice the rate.

There are disadvantages, however, other than the double sample rate. First, the aperture time of the sampling process must be small with respect to the period of the IF frequency. That is why a sample-and-hold circuit is shown in Fig. 11. If the A/D is flash, and the IF frequency is low enough, this circuit is not required. For IF frequencies as high as they are in the MAGR, certainly a silicon bipolar flash A/D would be required for this process. How quick the sample must be is debatable, other than it must be fast with respect to the IF frequency. Sin(x)/x, where x is proportional to the product of the frequency and the aperture time, attenuation occurs if the aperture time is too long, but then this attenuation also occurs on the noise, which is also at the IF frequency. Thus, there would be no loss in signal-to-noise ratio to a point.

The second disadvantage is minor. That is, as described above, the sample rate must be high enough so that there is no significant delay between the I and Q samples. This delay should be small with respect to a pseudonoise (PN) chip so that most of the time the I and Q samples do not straddle chip transitions. Thus, there is some loss associated with sampling right at twice the Nyquist frequency (four times the code chipping rate), but it is minimal. It has an effect similar to filter phase distortion.

This delay between the I and Q also has an effect on interference that may be present in the IF bandwidth. For example, consider an interference signal that is offset in frequency from the center of the IF band by a large amount, but is still in band. The sampling process also generates I and Q of the interference. The delay between these I and Q can be significant if the frequency offset is large, but, for normal receiver processing, this is acceptable. The effect is that some of the interference energy is folded over to the other side of center. Energy cannot be created, so this is of no consequence. Both sides of the center frequency are spread by the code correlation process.

As previously stated, the GPSCard™ uses IF sampling. Note from Fig. 8 that its IF frequency and sample frequency are such that N is 7 with a frequency offset of -1.00105 MHz. There are also at least 10 individual I and Q samples per chip, so the loss caused by the second disadvantage is negligible. Furthermore, the sample rate is not an integer multiple of the chipping rate, so that the I and Q sample times will never stay synchronous with the chip transitions. Because the IF frequency is relative low with a period of approximately 30 ns, the CMOS A/D aperture time poses no problem. Five levels of the A/D are used for a 2.5-bit quantization.

The MAGR performs baseband sampling, as shown in Fig. 10 using two 1.5-bit quantization (three levels: −L, 0, and +L) A/D converters. The result is the same as IF sampling, where

$$I_{sk} = \frac{A}{\sqrt{2}} C_k D_k \cos \phi_k \tag{24}$$

$$Q_{sk} = \frac{A}{\sqrt{2}} C_k D_k \sin \phi_k \tag{25}$$

with the exception that the Q sample is one sample later in the case of IF sampling. For the reasons stated above, we neglect that fact in the following discussions, although, in some applications, it is important.

The noise samples are simply as follows:

$$I_{nk} = x(t_k)/\sqrt{2} = x_k/\sqrt{2} \tag{26}$$

$$Q_{nk} = y(t_k)/\sqrt{2} = y_k/\sqrt{2} \tag{27}$$

The MAGR samples at a frequency at one-half of the 43.8 MHz clock from the PLL chip shown in Fig. 7, or 21.9 MHz. Given that this sampling is done at baseband, it is slightly more than the Nyquist sampling frequency, with at least two samples per P-code chip. Note that this sample rate is also not an integer multiple of the chipping rate, thus the sampling will never be synchronized with the chip transition times. The A/D threshold control is for CW interference suppression. This is a topic of discussion in Chapter 10, this volume.

2. *Precorrelation Filtering*

Precorrelation filtering (low-pass filters in Fig. 10 and bandpass filter in Fig. 11) is necessary to prevent aliasing while sampling in a digital receiver. However, this filtering also causes correlation losses, because the sidelobes of the PRN code spectrum are eliminated. Figure 12 illustrates this loss for a sharp cutoff,

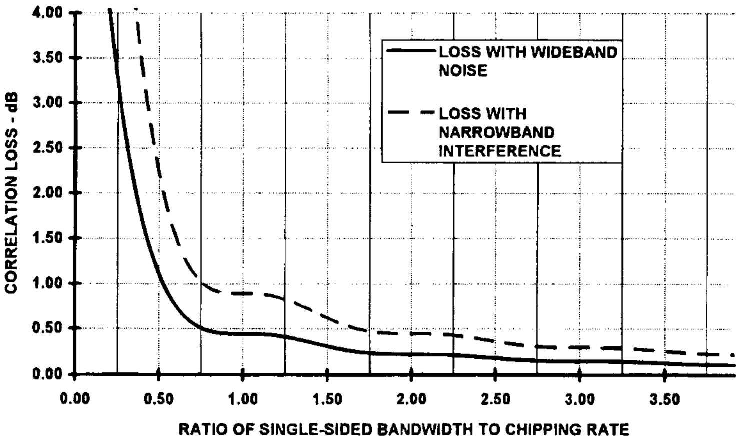


Fig. 12 Correlation loss caused by filtering.

Chapter 9

GPS Navigation Algorithms

P. Axelrad*
University of Colorado, Boulder, Colorado 80309
and
R. G. Brown†
Iowa State University, Ames, Iowa 50010

I. Introduction

THE previous chapters described the hardware and software needed to make GPS observations in the receiver. This chapter focuses on how these observations are processed to form a navigation solution. Fundamentally, a navigation solution is an estimate of the user position plus any other required parameters. The term "state" is used to describe all the parameters to be determined. The typical states in a GPS navigation estimator are three components of position, clock offset, and clock drift. In a moving application, three components of velocity are added. There are many applications described in the companion volume in which GPS is integrated with one or more other sensors, such as an altimeter or an inertial navigation system (INS). In such configurations, the state may be expanded to include specific sensor error states; however, in this chapter, we restrict ourselves to stand-alone GPS navigation estimation.

A navigation algorithm embedded in the GPS receiver combines raw measurements from the signal processor with GPS satellite orbit data to estimate the observer state. This process requires two sets of models—a measurement model and a dynamics or process model. The dynamics model describes the evolution of the system state. The measurement model relates the state to the GPS observations.

Section II describes the GPS measurements and Sec. III shows how they may be combined into a single point navigation estimate. Section IV provides various dynamic models used in GPS. The Kalman filter and some variations are described in Sec. V, and specific numerical examples are given in Sec. VI. For further information on filtering, the reader is advised to refer to Ref. 1, which specifically addresses GPS navigation filters, or more generally, Refs. 2, 3, or 4.

*Assistant Professor, Department of Aerospace Engineering Sciences.

†Distinguished Professor Emeritus, Department of Electrical and Computer Engineering.

II. Measurement Models

We consider three types of GPS measurements—pseudorange, Doppler, and accumulated delta range (ADR). The specifics of how a receiver actually forms these measurements is discussed in Chapters 7 and 8, this volume. Here we concentrate on the mathematical models of how the observations relate to the state of the vehicle.

A. Pseudorange

When the signal processor delay lock loop (DLL) finds the point of maximum correlation with a given GPS satellite signal, it produces an observation of the code phase, or equivalently, signal transmit time t_T for the current local receive time t_R. (In most cases, the observation is not a function of the filter navigation solution; however, the vector delay lock loop described in Chapter 8 of this volume provides a means to integrate the DLL and filter functions for improved performance.) The observed signal propagation delay is $(t_R - t_T)$. The pseudorange observable is merely this time interval scaled by the speed of light in a vacuum: $\rho = c\ (t_R - t_T)$.

The pseudorange observation between a user and satellite i can be related to the user position and clock states as follows:

$$\rho_i = |\boldsymbol{r}_i - \boldsymbol{r}_u| + c \cdot b_u + \epsilon_{\rho_i} \tag{1}$$

Where $\boldsymbol{r}_i$ is the satellite position at transmit time; $\boldsymbol{r}_u$ is the receiver position at receive time; b_u is the bias in the receiver clock (in s), and ϵ_ρ is the composite of errors produced by atmospheric delays, satellite ephemeris mismodeling, selective availability (SA), receiver noise, etc. (in m). Chapter 11, this volume provides an error budget for ϵ_ρ under various conditions.

The state to be estimated, consisting of $\boldsymbol{r}_u$ and $c \cdot b_u$ is embedded in this measurement equation. To extract it we must linearize the measurement equation about some nominal value, for example, about our current best estimate.

Given an a priori estimate of the state $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{r}}_u{}^T\ c \cdot \hat{b}_u]^T$ and an estimate of the bias contributions caused by ionospheric and tropospheric delay, relativistic effects, satellite clock errors $\hat{\epsilon}_{\rho_i}$, we can predict what the pseudorange measurement should be as follows:

$$\hat{\rho}_i = |\boldsymbol{r}_i - \hat{\boldsymbol{r}}_u| + c \cdot \hat{b}_u + \hat{\epsilon}_{\rho_i} \tag{2}$$

The measurement residual $\Delta\rho$, which is the difference between the predicted and actual measurement, can be modeled as linearly related to the error in the state estimate, $\Delta x \equiv [\Delta r^T\ c \cdot \Delta b]^T$, by performing a Taylor expansion about the current state estimate. The linearized result is given by the following:

$$\Delta\rho_i = \hat{\rho}_i - \rho_i = [-\hat{\mathbf{1}}_i^T \quad 1]\begin{bmatrix} \Delta\boldsymbol{r} \\ c \cdot \Delta b \end{bmatrix} + \Delta\epsilon_{\rho_i} \tag{3}$$

where

$$\hat{\mathbf{1}}_i \equiv \frac{\boldsymbol{r}_i - \hat{\boldsymbol{r}}_u}{|\boldsymbol{r}_i - \hat{\boldsymbol{r}}_u|}, \qquad \Delta\boldsymbol{r} \equiv \hat{\boldsymbol{r}}_u - \boldsymbol{r}_u, \qquad \Delta b \equiv \hat{b}_u - b_u, \qquad \Delta\epsilon_{\rho_i} \equiv \hat{\epsilon}_{\rho_i} - \epsilon_{\rho_i}$$

$\hat{\mathbf{1}}_i$ is the estimated line of sight unit vector from the user to the satellite; and

$\Delta\epsilon_{\rho_i}$ is the residual error after the known biases have been removed. This linearized model is the fundamental GPS pseudorange measurement equation.

The residual measurement error $\Delta\epsilon_{\rho_i}$ is generally composed of a slowly varying term, usually dominated by SA in civilian receivers, plus random or white noise. The expected variance of the error is required for any weighted navigation solution algorithm. An order of magnitude estimate of the slow terms can be obtained from the user equivalent range error (URE) reported in the Navigation message (see Chapter 4, this volume). The high-frequency error is produced primarily by receiver noise and quantization. For a typical receiver, the standard deviation is about 1/100 of the code chip, or about 3 m for C/A code and 0.3 m for P code. A more precise estimate can be based on the signal-to-noise ratio calculation in the channel, as described in Chapter 8, this volume.

B. Doppler

The numerically controlled oscillator (NCO), which controls the carrier-tracking loop, provides an indication of the observed frequency shift of the received signal. This observed frequency differs from the nominal L_1 or L_2 frequency because of Doppler shifts produced by the satellite and user motion, as well as the frequency error or drift of the satellite and user clocks. The Doppler shift caused by satellite and user motion is the projection of the relative velocities onto the line of sight scaled by the transmitted frequency $L_1 = 1575.42$ MHz divided by the speed of light, as follows:

$$D_i = -\left(\frac{\boldsymbol{v}_i - \boldsymbol{v}_u}{c}\cdot\mathbf{1}_i\right)L_1 \tag{4}$$

The Doppler can be converted to a pseudorange rate observation given by the following:

$$\dot{\rho}_i = (\boldsymbol{v}_i - \boldsymbol{v}_u)\cdot\frac{\boldsymbol{r}_i - \boldsymbol{r}_u}{|\boldsymbol{r}_i - \boldsymbol{r}_u|} + f + \epsilon_{\dot{\rho}_i} \tag{5}$$

where f is the receiver clock drift in m/s; and $\epsilon_{\dot{\rho}_i}$ is the error in the observation in m/s. Again, this effect can be predicted, based upon the current estimates of the velocity $\hat{\boldsymbol{v}}_u$; line of sight vector $\hat{\mathbf{1}}_i$, the clock drift estimate in m/s $\hat{f}$; and the known error rates $\hat{\epsilon}_{\dot{\rho}_i}$ as follows:

$$\hat{\dot{\rho}}_i = (\boldsymbol{v}_i - \boldsymbol{v}_u)\cdot\hat{\mathbf{1}}_i + \hat{f} + \hat{\epsilon}_{\dot{\rho}_i} \tag{6}$$

The linearized Doppler measurement equation is then as follows:

$$\Delta\dot{\rho}_i \equiv \hat{\dot{\rho}}_i - \dot{\rho}_i = [-\hat{\mathbf{1}}_i^T \quad 1]\begin{bmatrix}\Delta \boldsymbol{v}\\ \Delta f\end{bmatrix} + \Delta\epsilon_{\dot{\rho}_i} \tag{7}$$

Note that the Doppler does depend on the observer position through the line-of-sight unit vector. This dependence can be exploited to perform "Doppler positioning" in which the position is solved for using the Doppler observations and, sometimes, the rate of change of Doppler. This is the positioning method employed with the Navy's Transit satellites. The observation geometry for Doppler positioning is substantially weaker than ranging; thus, it is not used much in GPS, except to set an a priori position estimate.

C. Accumulated Delta Range

The ADR is produced by the signal processor by accumulating the commanded values to the NCO required to maintain lock on the signal. In other words, it keeps track of changes in the observed range to the satellite. Thus, both terms "accumulated delta range" or "integrated Doppler" are appropriate. In the literature (cf. Ref. 5) this measurement has also been called "carrier beat phase," referring to the output of a mixing process between a nominal L_1 carrier signal generated in the receiver, and the received Doppler-shifted version.

The distinction between the ADR and a code-based pseudorange is that the ADR has an ambiguous starting value. Once the phase-lock-loop (PLL) begins to follow the carrier signal, it can keep track of the total change in range; however, there is no way to know the whole number of carrier cycles between the satellite and the user antenna. Thus, for stand-alone navigation it is not possible to use the ADR for absolute estimation of position. (The ADR initial condition problem is similar to that encountered in inertial navigation systems.) However, for differential GPS or attitude determination, it is possible to determine the difference in the ambiguities between two nearby stations and/or two GPS satellites. Thus, for these applications, the precision of the ADR can be fully exploited.

A common use of the ADR in stand-alone navigation is to smooth the noisy pseudorange measurements. A number of techniques are available for doing this. A commonly used technique forms a weighted average of the code and carrier-based measurements.[6] Reference 7 has also suggested an integration scheme where the ADR provides the reference trajectory (in much the same manner as an INS), and then the pseudorange data is used at a slower rate to update the reference trajectory via a Kalman filter. Other ad hoc methods have also been successfully employed for real-time, stand-alone applications. The ADR measurements play only a minor role in many GPS stand-alone navigation applications; however, they are pivotal to kinematic differential operations and surveying, as described in Chapters 15 and 18 in the companion volume.

D. Navigation Data Inputs

To compute the predicted pseudorange and Doppler, the navigation algorithm must have information on the position and velocity of the GPS satellite, as well as error models to correct the satellite clock offset and atmospheric delays. This information is provided via the Navigation message. The satellite positions are computed as described in Chapter 4, this volume; the atmospheric delay models are described in Chapter 8, this volume; and the clock corrections are described in Chapter 4, this volume. This information is also contained in Ref. 8.

III. Single-Point Solution

The physical measurements and equations provided in the previous section are all that is required for a single-point solution or kinematic solution. In this method, the navigation estimate is the least squares solution to the measurement equations made at a single time. For each satellite tracked by the receiver, the predicted pseudorange is formed using Eq. (2), and the linearized observation Eq. (3) is formed. All the measurements are then combined into a set of normal equations:

$$\Delta\rho = G\Delta x + \Delta\epsilon_\rho \tag{8}$$

where

$$\Delta\rho \equiv \begin{bmatrix} \Delta\rho_1 \\ \Delta\rho_2 \\ \vdots \\ \Delta\rho_n \end{bmatrix}, \quad G \equiv \begin{bmatrix} -\hat{\mathbf{1}}_1^T & 1 \\ -\hat{\mathbf{1}}_2^T & 1 \\ \vdots & \vdots \\ -\hat{\mathbf{1}}_n^T & 1 \end{bmatrix}, \quad \Delta \boldsymbol{x} \equiv \begin{bmatrix} \Delta \boldsymbol{r}_u \\ c \cdot \Delta b_u \end{bmatrix}, \quad \Delta\epsilon_\rho \equiv \begin{bmatrix} \Delta\epsilon_{\rho_1} \\ \Delta\epsilon_{\rho_2} \\ \vdots \\ \Delta\epsilon_{\rho_n} \end{bmatrix}$$

which is to be solved for a correction, $\Delta \boldsymbol{x}$ to the a priori state estimate. To improve the state estimate subtract $\Delta \boldsymbol{x}$ from the a priori values. In GPS, "G" is frequently referred to as the geometry matrix, and corresponds to the measurement connection matrix, commonly named "H" in the more general literature on filtering.

The $\Delta\epsilon_\rho$ are assumed to be zero mean, so that the least squares solution to the set of normal equations is given by the following:

$$\Delta\hat{\boldsymbol{x}} \equiv (G^TG)^{-1}G^T\Delta\rho \tag{9}$$

or, if a weight R_i^{-1} is assigned to each observation, the weighted least squares estimate is as follows:

$$\Delta\hat{\boldsymbol{x}} \equiv (G^TR^{-1}G)^{-1}G^TR^{-1}\Delta\rho \tag{10}$$

If the a priori estimate used to construct G, is off by a lot (typically more than a few km), the least squares solution may be iterated until the change in the estimate is sufficiently small. Because G only depends upon the line-of-sight unit vector, it is not very sensitive to errors in the observer position. Numerically efficient methods for solving Eqs. (8–10) are well known (c.f. Ref. 9).

A. Solution Accuracy and Dilution of Precision

How accurate is the single-point, least squares solution? The accuracy is decided by two factors, the measurement quality and the user-to-satellite geometry. [Chapter 11, this volume, on errors in GPS details the contributions of both measurement errors and geometry, and Chapter 5, this volume, provides an extensive discussion of geometric dilution of precision (GDOP).] The measurement quality is described by the variance of the measurement error, which for a typical pseudorange is in the range of 0.3 to 30 m, depending on the error conditions. The geometry is described by the "G" matrix, which is composed of line-of-sight vectors and "1s" for the clock states.

The solution error covariance can be expressed as follows:

$$\begin{aligned} E[\Delta\hat{x} \quad \Delta\hat{x}^T] &= E[(G^TG)^{-1}G^T\Delta\rho \quad \Delta\rho^T G(G^TG)^{-1}] \\ &= (G^TG)^{-1}G^TRG(G^TG)^{-1} \end{aligned} \tag{11}$$

where R is the pseudorange measurement covariance. If we assume (somewhat incorrectly, as described later) that the measurement errors are uncorrelated and have equal variance σ^2 then $R = \sigma^2 I$, and the point solution error covariance reduces to the following:

$$E[\Delta\hat{x} \quad \Delta\hat{x}^T] = \sigma^2(G^TG)^{-1} \tag{12}$$

If the state is parameterized so that $\Delta x = [\Delta E\ \Delta N\ \Delta U\ c \cdot \Delta b]^T$, where ΔE, ΔN, and ΔU, are the east, north, and up position errors, respectively; and $c \cdot \Delta b$ is the clock bias error, then the variance of the state estimates is given by the following:

$$E[\Delta\hat{x}\ \Delta\hat{x}^T] = \begin{bmatrix} E[\Delta E^2] & E[\Delta E\Delta N] & E[\Delta E\Delta U] & E[\Delta Ec\cdot\Delta b] \\ E[\Delta N\Delta E] & E[\Delta N^2] & E[\Delta N\Delta U] & E[\Delta Nc\cdot\Delta b] \\ E[\Delta U\Delta E] & E[\Delta U\Delta N] & E[\Delta U^2] & E[\Delta Uc\cdot\Delta b] \\ E[c\cdot\Delta b\Delta E] & E[c\cdot\Delta b\Delta N] & E[c\cdot\Delta b\Delta U] & E[c\cdot\Delta b^2] \end{bmatrix} \tag{13}$$

Most of the time, we are primarily interested in the diagonal elements. The following DOPs summarize the contribution of the geometry:

$$\begin{aligned} A &\equiv (G^TG)^{-1} \\ \text{GDOP} &\equiv \sqrt{\text{trace}(A)} && \text{geometrical DOP} \\ \text{PDOP} &\equiv \sqrt{A_{11} + A_{22} + A_{33}} && \text{position DOP} \\ \text{HDOP} &\equiv \sqrt{A_{11} + A_{22}} && \text{horizontal DOP} \\ \text{VDOP} &\equiv \sqrt{A_{33}} && \text{vertical DOP} \\ \text{TDOP} &\equiv \sqrt{A_{44}} && \text{time DOP} \end{aligned} \tag{14}$$

Thus, the total position error magnitude can be estimated by $\sigma \times$ PDOP, and the vertical position error by $\sigma \times$ VDOP, etc. However, keep in mind that this is only an approximation, because of the assumption that all satellite pseudorange measurements errors are independent and have the same statistics. The equations for a single-point velocity solution are identical with the pseudorange measurements replaced by pseudorange rates.

Typical modern receivers track 5–12 satellites simultaneously. More satellites produce improved geometry, generally leading to a more accurate single-point navigation solution. (An exception can occur if the ranging error to the additional satellite is exceptionally poor.) ADR smoothing can reduce the receiver-induced measurement noise in each observation; however, it cannot eliminate the effects of SA or atmospheric effects. To improve the navigation estimate further, we must tie together measurements over time by including knowledge of the vehicle dynamics in a solution filter.

B. Point Solution Example

As an example of the least squares solution method, assume an observer is actually located on the surface of the Earth at 0° latitude, 0° longitude, and has a clock error from GPS time equivalent to 85,491.5 m. If the observer state is comprised of the WGS-84 (1984 Word Geodetic System) position components x, y, z, and the clock bias in meters, the true state is as follows:

$$\boldsymbol{x} = [6{,}378{,}137.0\ \text{m} \quad 0.0\ \text{m} \quad 0.0\ \text{m} \quad 85{,}000.0\ \text{m}]^T$$

At a certain time there are seven satellites visible above an elevation of 10 deg at the positions shown in Table 1.

If the a priori position and clock estimate is given by

$$\hat{\boldsymbol{x}} = [6{,}377{,}000.0\ \text{m} \quad 3{,}000.0\ \text{m} \quad 4{,}000.0\ \text{m} \quad 0.0\ \text{m}]^T$$

then the computed range and line-of-sight unit vector to each satellite are as shown in Table 2.

Table 1 Satellite positions for point solution example

Satellite	X position, m	Y position, m	Z position, m
SV 01	22,808,160.9	−12,005,866.6	−6,609,526.5
SV 02	21,141,179.5	−2,355,056.3	−15,985,716.1
SV 08	20,438,959.3	−4,238,967.1	16,502,090.2
SV 14	18,432,296.2	−18,613,382.5	−4,672,400.8
SV 17	21,772,117.8	13,773,269.7	6,656,636.4
SV 23	15,561,523.9	3,469,098.6	−21,303,596.2
SV 24	13,773,316.6	15,929,331.4	−16,266,254.4

Table 2 Computed pseudorange and line-of-sight vectors

Satellite	Computed pseudorange, m	Line-of-sight X	Line-of-sight Y	Line-of-sight Z
SV 01	21,399,408.0	0.767832	−0.561178	−0.309052
SV 02	21,890,921.6	0.674443	−0.107718	−0.730427
SV 08	22,088,910.4	0.636607	−0.192041	0.746895
SV 14	22,666,464.0	0.531856	−0.821318	−0.206314
SV 17	21,699,943.6	0.709454	0.634576	0.306574
SV 23	23,460,242.4	0.391493	0.147744	−0.908243
SV 24	23,938,978.9	0.308965	0.665289	−0.679655

The computed geometry matrix is as follows:

$$G = \begin{bmatrix} -0.767832 & 0.561178 & 0.309052 & 1 \\ -0.674443 & 0.107718 & 0.730427 & 1 \\ -0.636607 & 0.192041 & -0.746895 & 1 \\ -0.531856 & 0.821318 & 0.206314 & 1 \\ -0.709454 & -0.634576 & -0.306574 & 1 \\ -0.391493 & -0.147744 & 0.908243 & 1 \\ -0.308965 & -0.665289 & 0.679655 & 1 \end{bmatrix}$$

Table 3 provides the simulated measured pseudorange (already corrected for known errors such as ionospheric delay, satellite clock, etc.) and the pseudorange residual $\Delta\rho_i \equiv \hat{\rho}_i - \rho_i$. The standard deviation of the pseudorange errors is 6 m. Note that the pseudorange residuals are dominated by the large error in the estimate of the receiver clock.

Solving the normal equations for a correction to the state estimate gives $\Delta x = [-1{,}131.8 \quad 2{,}996.8 \quad 3{,}993.1 \quad -84{,}996.4]^T$ m. Subtracting this from the a priori estimate gives the improved estimate $\hat{x} = [6{,}378{,}131.8 \quad 3.2 \quad 6.9 \quad 84{,}996.4]^T$ m. This new estimate is closer to the true value of the state; however, it contains errors produced by the pseudorange measurement error as well asthe approximation in the line-of-sight vectors caused by the incorrect a priori guess. To see how large the latter effect is, we can redo the least squares solution using the improved estimate to compute the elements of the G matrix. The resulting correction to the state estimate is $\Delta x = [0.3 \quad -0.1 \quad -0.2 \quad 0.6]^T$ m, and the "improved" state estimate (actually worse than the last estimate) is $\hat{x} = [6{,}378{,}131.5 \quad 3.3 \quad 7.1 \quad 84{,}995.8]^T$ m. Thus, it is apparent that the approximation made in computing the G matrix was quite good, and to get a solution at the 1-m level, it is not necessary to iterate if the solution is already known to within a few kilometers.

The final error in the state estimate is $\Delta x = [-5.5 \quad 3.2 \quad 7.1 \quad -4.2]^T$ m. Now let us compare this to the error bound predicted by the GDOP approximation. The A matrix can be computed from the G matrix given in Eq. (14). The DOPS for each of the state components are computed as the square roots of the diagonal elements of A. The measurement standard deviation $\sigma = 6$ m.
Thus, the DOP approximations seem to be valid for this example.

Table 3 Simulated pseudorange and residual

Satellite	Measured pseudorange, m	Pseudorange residual, m
SV 01	21,480,623.2	−81,215.3
SV 02	21,971,919.2	−80,997.6
SV 08	22,175,603.9	−86,693.4
SV 14	22,747,561.5	−81,097.6
SV 17	21,787,252.3	−87,308.8
SV 23	23,541,613.4	−81,371.0
SV 24	24,022,907.4	−83,928.6

Table 4 Actual point solution errors compared to DOP predictions

State component	DOP value	Expected $1 - \sigma$ error, $\sigma \times$ DOP	Actual error
X position	3.0	18.0 m	−5.5 m
Y position	0.8	4.8 m	3.2 m
Z position	0.8	4.8 m	7.1 m
Clock bias	1.9	11.4 m	−4.2 m
Total error	3.7	22.2 m	10.4 m

IV. User Process Models

In anticipation of employing Kalman filter methods in the GPS solution (Sec. V), we now look at user "process models." The vehicle dynamics are summarized in the filter process model. The GPS has the capability to provide real-time three-dimensional position, velocity, and time information to any user. However, there are times when all this information is not required or valuable. The degree to which the user dynamics are constrained or predictable dictates the type of process model used.

A. Clock Model

Two states required in any GPS-based navigation estimator are the user clock bias and drift, which represent the phase and frequency errors in the atomic frequency standard or crystal oscillator in the receiver. Within the navigation algorithm the two-state model shown in Fig. 1 is commonly employed.

This model says that we expect both the frequency and phase to random walk over a short period of time. The discrete process equations are given by Ref. 1.

$$\boldsymbol{x}_c(k) = \Phi_c(\Delta t)\boldsymbol{x}_c(k-1) + \boldsymbol{w}_c(k-1) \tag{15}$$

where

$$\boldsymbol{x}_c \equiv \begin{bmatrix} b \\ f \end{bmatrix}, \qquad \Phi_c(\Delta t) = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}$$

$$Q_c \equiv E[\boldsymbol{w}_c \boldsymbol{w}_c^T] = \begin{bmatrix} s_b\Delta t + S_f\dfrac{\Delta t^3}{3} & S_f\dfrac{\Delta t^2}{2} \\ S_f\dfrac{\Delta t^2}{2} & S_f\Delta t \end{bmatrix}$$

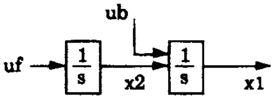


Fig. 1 Dynamic model for GPS clock states.

The white noise spectral amplitudes S_b and S_f can be related to the classical Allan variance parameters. The approximate relation given in Ref. 1 (p. 427) is $S_f = 2\,h_0$, and $S_g = 8\,\pi^2\,h_{-2}$ (see also Ref. 10). Figure 2 shows simulated clock states for a crystal oscillator with $h_0 = 2 \times 10^{-19}$, and $h_{-2} = 2 \times 10^{-20}$ (Ref. 1, Chapter 10).

Two clock states of this type must be included for all types of GPS users. In a time transfer receiver, which is described in greater detail in Chapter 16 of the companion volume, these two clock states are the ones of primary interest. In this case, the receiver position is generally known to at least the level of accuracy of timing information desired (1 m ~ 3 ns). For highest accuracy, the position is held fixed and only the two clock parameters are estimated.

B. Stationary User or Vehicle

If the user antenna is known to be stationary at an unknown location, three position coordinate states may be added to the clock model to form a 5-element state vector. It is assumed that the velocity is zero, thus, the dynamic model for

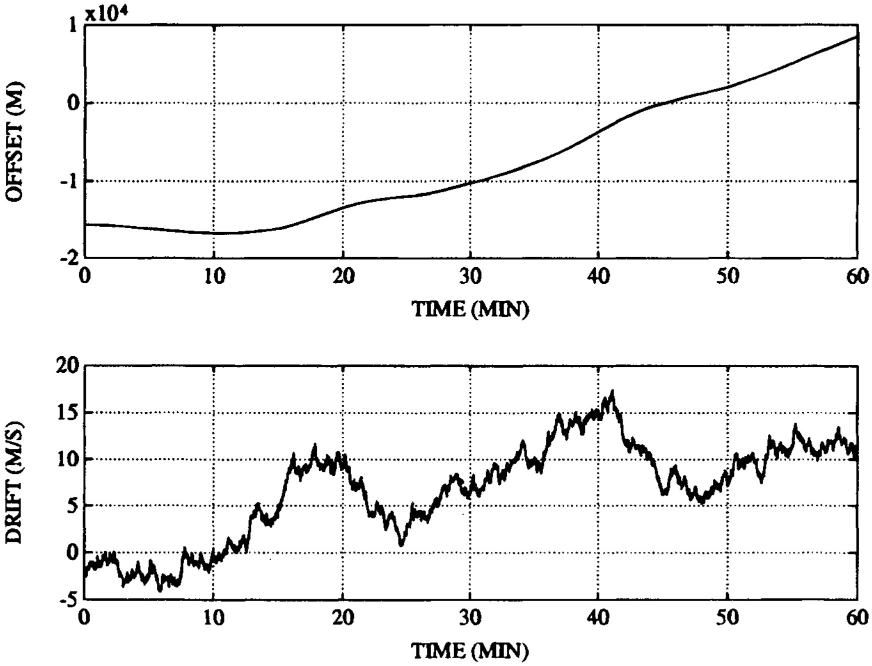


Fig. 2 Simulated GPS clock errors. The bottom graph shows an example of clock frequency drift measured in m/s. The top graph shows the corresponding clock offset in meters. Note, the receiver clock errors can be very large, in this case more than 10 km, and must be estimated along with the position solution.

the stationary user is given by the following:

$$\boldsymbol{x}_s(k) = \Phi_s(\Delta t)\boldsymbol{x}_s(k-1) + \boldsymbol{w}_s(k-1) \tag{16}$$

where

$$\boldsymbol{x}_s \equiv [x \quad y \quad z \,|\, b \quad f]^T$$

$$\Phi_s(\Delta t) = \left[\begin{array}{c:c} I & 0 \\ \hdashline 0 & \Phi_c(\Delta t) \end{array}\right], \qquad \text{and } I \text{ is a } 3 \times 3 \text{ identity matrix}$$

$$Q_s \equiv E[\boldsymbol{w}_s \boldsymbol{w}_s^T] = \left[\begin{array}{c:c} Q_p & 0 \\ \hdashline 0 & Q_c \end{array}\right]$$

There are two important things to note. First, it cannot be assumed that the clock state is constant; thus, the frequency error state is required as well as the bias. Second, even for a stationary result, we model the dynamics by a random walk to prevent numerical problems in the navigation algorithm. The process noise covariance Q_p represents the uncertainty in the dynamic model. Thus, for a stationary observer, we would think that it could be set to zero. This is not generally done because it can lead to numerical problems or cause the filter to "go to sleep." In this situation, the estimation error covariance has decreased so far that the estimator gain for new measurements goes to zero—essentially the filter begins to ignore new information. As long as this situation is avoided, Q_p can be set to a small value to maximize the smoothing that will occur.

C. Low Dynamics

The next step up in user dynamics is a low dynamic vehicle, such as a boat or car. In these cases, the position, velocity, and clock terms must be estimated, leading to an 8-element state representation. The discrete model for such a user is shown in Fig. 3. The corresponding dynamic model is given by the following:

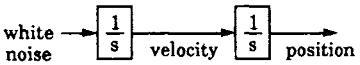


Fig. 3 Integrated random-walk model for a dynamic observer.

$$\boldsymbol{x}_L(k) = \Phi_L(\Delta t)\boldsymbol{x}_L(k-1) + \boldsymbol{w}_L(k-1) \tag{17}$$

where

$$\boldsymbol{x}_L \equiv [x \quad y \quad z \,|\, \dot{x} \quad \dot{y} \quad \dot{z} \,|\, b \quad f]^T$$

$$\Phi_L(\Delta t) = \left[\begin{array}{cc:c} I & \Delta t I & 0 \\ 0 & I & 0 \\ \hdashline 0 & 0 & \Phi_c(\Delta t) \end{array}\right]$$

$$Q_L \equiv E[\boldsymbol{w}_L \boldsymbol{w}_L^T] = \left[\begin{array}{cc:c} Q_p & Q_{pv} & 0 \\ Q_{pv} & Q_v & 0 \\ \hdashline 0 & 0 & Q_c \end{array}\right]$$

The effect of unknown random accelerations between measurement updates is represented by Q_υ. Often, different values are used for horizontal and vertical components; i.e., a car cannot change its vertical velocity substantially; whereas, it can accelerate or decelerate rapidly. If the dynamical uncertainty of the vehicle is large, filtering will not improve the navigation solution.

D. High Dynamics

When the vehicle, such as a fighter aircraft or missile, has the potential for significant accelerations, it is usually necessary to measure and account for the deterministic changes in velocity. This leads to the integrated GPS/INS system, which is discussed in detail in Chapter 2 of the companion volume. A less accurate way of handling the high dynamics problem is to add three acceleration states to the process model and let the stand-alone GPS system estimate the vehicle acceleration in addition to position and velocity. The acceleration states are usually modeled as either random walk or Markov processes. This method of coping with high dynamics is not as good as a full-fledged integrated GPS/INS system, but it is better than treating acceleration as white noise, which is what has to be done if the acceleration states are omitted.

V. Kalman Filter and Alternatives

As mentioned previously, one disadvantage of the point solution approach is that it does not carry any information from one measurement epoch to the next; i.e. it does not include any of the known user dynamics. A second problem is that the solution accuracy is extremely dependent on the instantaneous satellite geometry. Employing a Kalman filter addresses both of these issues (Ref. 1).

A. Discrete Extended Kalman Filter Formulation

For use with stand-alone GPS, a discrete, extended Kalman filter (EKF) is generally used. This means that measurements are incorporated at discrete intervals, and the measurement models are linearized about the current best estimate of the state. The updated state estimate is formed as a linear blend of the previous estimate (projected forward to the current time) and the current measurement information. The relative weighting in the blend is determined by the a priori error covariance and the measurement error covariance. After updating, the state estimate and its error covariance matrix are projected ahead to the next measurement time via the assumed process dynamics.

The state x includes the two clock components, three position components, and possibly three velocity components, depending on the type of dynamic model used. Associated with the state are four key matrices that must be specified in the discrete EKF:

G—the measurement connection matrix, the elements of which are the partials of the measurement model with respect to each of the states (In most literature on filtering, this matrix is refered to as H; we use G for geometry matrix to be consistent with the GPS literature.)

R—the measurement noise covariance matrix

Φ—the state transition matrix, which is a linearized representation of the process model

Q—the process noise covariance matrix

The selection of R and Q has a significant effect on the convergence and accuracy of the filter solutions. The adjustment of these parameters is refered to as "filter tuning." Tuning is often performed to achieve the best possible performance while avoiding filter divergence in the face of unmodeled errors.

In addition to specifying the four matrices, we must also establish an initial estimate of the state $\hat{x}_0^-$ and the state covariance matrix ($P_0{}^-$). The filter proceeds by processing all available measurements at each epoch (the measurement update) and then propagating the state estimate and covariance ahead to the next epoch (the time update). Estimates of the state and covariance after the measurement update are indicated by a superscript "+"; estimates and covariances propagated ahead are indicated by a superscript "−".

The measurement update is summarized as follows:

1) Compute the expected pseudorange $\hat{\rho}_k$ according to Eq. (2) based on the GPS satellite position and the a priori state estimate $\hat{\boldsymbol{x}}_k^-$.

$$\hat{\rho}_k = h(\hat{\boldsymbol{x}}_k^-)$$

2) Construct the measurement connection matrix as follows:

$$G_k(\hat{\boldsymbol{x}}_k^-) \equiv \frac{\partial h(\hat{\boldsymbol{x}}_k^-)}{\partial \boldsymbol{x}}$$

3) Compute the gain matrix K_k according to the following:

$$K_k = P_k^- G_k^T (G_k P_k^- G_k^T + R_k)^{-1} \tag{18}$$

4) Update the state as follows:

$$\hat{\boldsymbol{x}}_k^+ = \hat{\boldsymbol{x}}_k^- + K_k(\rho_{k-\text{measured}} - \hat{\rho}_k) \tag{19}$$

(Note that the quantity in parenthesis is opposite in sign to the measurement residual $\Delta\rho = \hat{\rho}_k - \rho_k$ defined earlier.)

5) and the following covariance matrix*:

$$P_k^+ = (I - K_kG_k)P_k^-(I - K_kG_k)^T + K_kR_kK_k^T \tag{20}$$

The projection steps are as follows:

1) Propagate the covariance matrix to the next measurement epoch as follows:

$$P_{k+1}^- = \Phi_kP_k^+\Phi_k^T + Q_k \tag{21}$$

2) Propagate the state estimate to the next measurement epoch using the assumed process dynamics. If the dynamic model is linear, the propagation is given by the following:

$$\hat{\boldsymbol{x}}_{k+1}^- = \Phi_k\hat{x}_k^+ \tag{22}$$

In the more general case where the zero-noise dynamic model is given by the following:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x} = f(\boldsymbol{x}, t) \tag{23}$$

$\hat{\boldsymbol{x}}_k^+$ is projected forward by numerically integrating the nonlinear dynamic model. These steps are illustrated in Fig. 4.

The standard EKF assumes that the measurement and process noise are not correlated with each other and that each is uncorrelated between time epochs. For GPS, the former assumption is largely valid, but the latter is not. In the next section, methods for dealing with correlated measurement noise are discussed.

B. Steady-State Filter Performance

A key feature of the Kalman filter is that, under many conditions, it quickly converges to a quasi-steady-state condition. In GPS navigation, for example, the geometry matrix G changes rather slowly, and in all of the models considered here, the state transition matrix is only a function of the measurement time interval. Thus, after a short time, the increase in the error covariance caused by state propagation and dynamic uncertainty is matched by the decrease in error covariance caused by the measurement update. The discrete, steady-state Kalman filter can be derived by assuming that both P^+ and P^- are constants in Eqs. (18), (20), and (21). Substitution results in a discrete time Riccati equation that does converge but in general does not afford a closed form solution. This is described in further detail in Refs. 11 and 12. An interesting point is that in GPS, it frequently takes only a few measurement epochs to achieve this steady-state value.

* The form of the expression given in Eq. (20) for the covariance measurement update ensures that if P^- is symetric, P^+ will also be symetric. The more commonly used form $P_k^+ = (I - K_kH_k)P_k^-$, does not have this property. In this form, numerical difficulties can result, for example, if the initial covariance is very large and the measurements are very accurate.

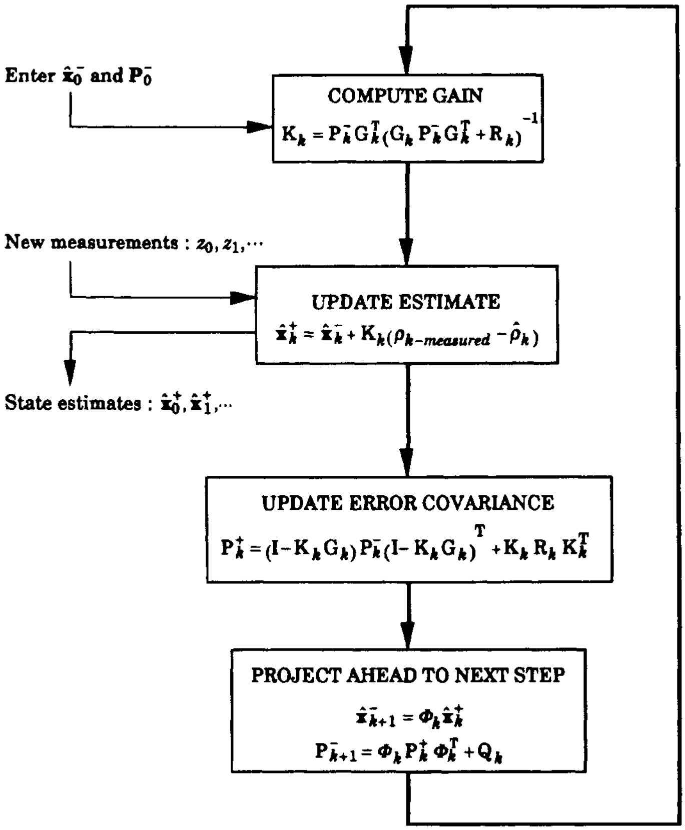

Enter $\hat{x}_0^-$ and P_0^-
COMPUTE GAIN
$K_k = P_k^- G_k^T (G_k P_k^- G_k^T + R_k)^{-1}$
New measurements : $z_0, z_1, \cdots$
UPDATE ESTIMATE
$\hat{x}_k^+ = \hat{x}_k^- + K_k(\rho_{k\text{-}measured} - \hat{\rho}_k)$
State estimates : $\hat{x}_0^+, \hat{x}_1^+, \cdots$
UPDATE ERROR COVARIANCE
$P_k^+ = (I - K_k G_k) P_k^- (I - K_k G_k)^T + K_k R_k K_k^T$
PROJECT AHEAD TO NEXT STEP
$\hat{x}_{k+1}^- = \Phi_k \hat{x}_k^+$
$P_{k+1}^- = \Phi_k P_k^+ \Phi_k^T + Q_k$

Fig. 4 Flow chart of the discrete Kalman filter solution.

C. Alternate Forms of the Kalman Filter

Under special conditions, there may be computational difficulties with the standard Kalman filter formulation shown above. These include numerical instability and divergence produced by round-off errors even when high precision arithmetic is employed. Square-root filters can mitigate these problems to some extent. Perhaps the most widely used square-root filter in GPS is the *U-D* filter. This is described in detail in Ref. 13, and in less detail in Refs. 1 and 4.

In the U-D filter, the covariance matrix P is decomposed into the factored form $P = UDU^T$. The measurement and time updates are formulated directly in terms of the U and D matrices, and conversion to the covariance form is only required for input and output purposes. In brief, the divergence problems encountered in propagating the U and D separately are less severe than they would be if P were propagated in unfactored form.

D. Dual-Rate Filter

A second implementation issue relates to the rate at which the various filter matrices are updated. The computational load of updating the measurement connection matrix, the covariance, and the Kalman gain at each measurement epoch may be a severe burden on the receiver processor. In reality, the measurement geometry changes rather slowly, and once the filter has converged, the covariance matrix tends to a steady value until there is a change in the satellites tracked. Thus, some GPS receivers employ a dual rate or background scheme for updating some of these parameters. This approach may have benefit in reducing the real-time processing load, but also has the potential to increase the software complexity.

E. Correlated Measurement Noise

One of the major difficulties for GPS navigation filters in civilian receivers is selective availability. Recall that the Kalman filter assumes that successive measurement errors are uncorrelated to each other; however, when SA is present, there is an unknown, slowly varying error associated with each satellite, as described in Chapter 16, this volume. Even without SA, measurement errors produced by orbit errors, ionosphere, troposphere, and multipath have time-correlated statistics. This clearly violates one of the basic filter assumptions; thus, even if the measurement noise as described by the R matrix represents the typical error variance, it does not correctly model the effect.

One approach to dealing with this problem would be to add a random walk state to the filter for each satellite tracked. However, for a dynamic user, there is not sufficient information to separate the SA from the vehicle motion effectively, nor is there any assurance that random walk (or any other one-state process) will properly model the SA process in effect at the moment.

Another method for dealing with this unknown signal dynamics is to include a "consider" state for each satellite in a Schmidt filter (Ref. 1, Chap. 9). The Schmidt filter accounts for the covariance of these additional states without trying to estimate the actual parameter values. This improves the overall state covariance estimate and leads to more realistic error bounds and better performance under SA. Of course, the Schmidt filter cannot remove the effect of SA; it does, however, account for it to within the bounds of our uncertain knowledge of SA.

Several methods are described in the literature that can be applied to correlated measurement errors for which the dynamic process governing the errors is known (cf. Refs. 1, 3, and 12). For example, an error that can be modeled by a first order Markov process with known variance and time constant can be readily accomodated in the filter. The key is to construct a new measurement that differences prior measurements to remove the correlated error. This results in modifications to the Kalman filter gain and covariance update equations, which are given in the references mentioned.

VI. GPS Filtering Examples

The following examples illustrate the set up and performance of a Kalman filter under a variety of conditions. All of the results were generated based on a simulation written in MATLAB® that includes the full primary GPS satellite constellation. The first case considered is a stationary buoy with no SA. The second is a vehicle traveling at constant velocity. The last two cases illustrate the effects of unmodeled dynamics and correlated measurement errors on the filter performance.

A. Buoy Example

As a first example of the operation of a Kalman Filter for GPS navigation, we model a fixed buoy located near the surface of the Earth at 0° latitude and 0° longitude. The buoy is nominally not moving, so the stationary dynamic model is used. A relatively large dynamic uncertainty in each position component of $Q_p(i,i) = (10\ \text{cm})^2$ over a 1-s interval is allowed to account for the possibility of random buoy motion on the water. The dynamic model for the clock is the second-order system described in Sec. IV.

Thus, the state is comprised of the three WGS-84 coordinates plus clock bias and drift components $\boldsymbol{x} = [x\ y\ z\ b\ f]^T$. The initial estimate of the state is

$$\hat{\boldsymbol{x}} = [6{,}377{,}000.0\ \text{m} \quad 3{,}000.0\ \text{m} \quad 4{,}000.0\ \text{m} \quad 0.0\ \text{m} \quad 0.0\ \text{m/s}]^T$$

The diagonal elements of the initial covariance matrix are selected to reflect the uncertainty in the a priori state estimates. Generally these can be set very large, because after the first set of measurements is received, the uncertainty will be reduced to approximately the level of the single-point solution. One caveat is that if there are large differences between the initial uncertainties in the states, it can lead to numerical problems if the alternate form of Eq. (20) is used. For this example we use the following:

$$\boldsymbol{P}_0^- = \begin{bmatrix} 25\ \text{km}^2 & & & & 0 \\ & 25\ \text{km}^2 & & & \\ & & 25\ \text{km}^2 & & \\ & & & 10^4\ \text{km}^2 & \\ 0 & & & & \left(c \cdot 10^{-6}\,\dfrac{\text{km}}{\text{s}}\right)^2 \end{bmatrix}$$

The receiver takes measurements to all GPS satellites at an elevation above 10 deg, at intervals of 1 s. Each pseudorange measurement is assumed to have a variance $R = 36\ \text{m}^2$.

Figure 5 presents the estimation errors for the $\boldsymbol{x}$ position and clock bias states over a 20 s simulation run. The graphs show the actual error in the estimate as well as the $\pm 1 - \sigma$ bounds computed from the diagonals of the filter covariance matrix. (The initial a priori covariance is off the scale.) This clearly illustrates the convergence of the filter. As each set of measurements is incorporated into the filter, the state uncertainty is decreased by an amount related to the a priori covariance, the measurement variance, and the measurement geometry. When the state is projected ahead to the next measurement epoch, the uncertainty increases because of the limitations of our knowledge of the governing dynamics.

After only about five epochs, the Kalman filter reaches the almost steady-state condition in which the amount of information gained in the measurement step is equal to the loss of information in the projection step. If the satellite geometry

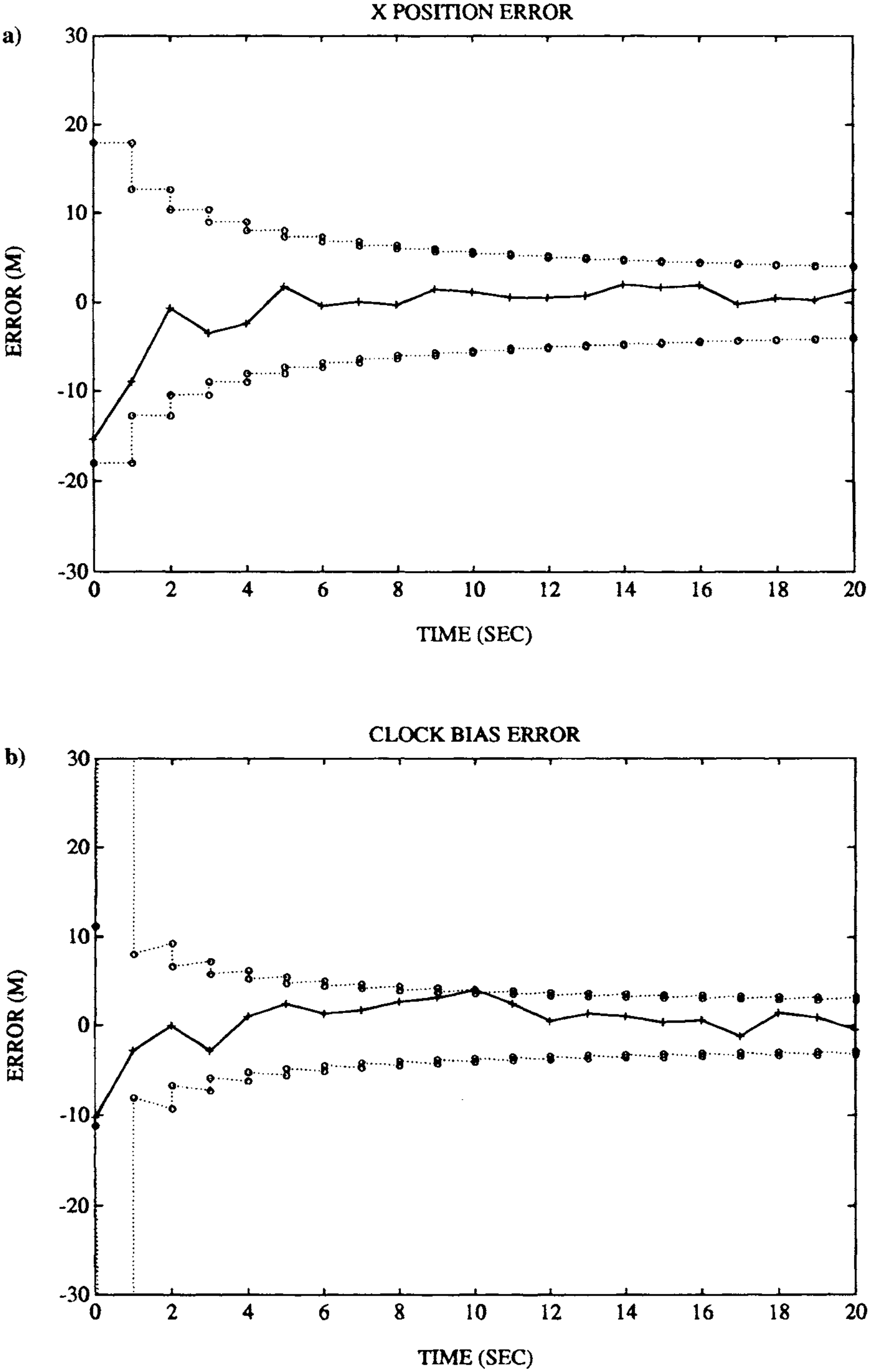


Fig. 5 Kalman filter estimation errors for buoy example. In both graphs the solid line with the "+" symbols represents the actual error. The dotted lines with the "o" symbols represent the filter computed 1 − σ bounds both before and after the measurement update; a) *x* position estimation error; b) clock bias estimation error.

remains fairly consistent, this condition will continue. If the geometry degrades or if measurements are missed, the uncertainty will grow at a rate determined by the process noise matrix.

For comparison with the Kalman filter results, Fig. 6 shows the errors in the z position and clock bias point solutions utilizing the same measurement data. Note that although the filter solution uncertainty was reduced to less than 5 m within 20 s, the point solution accuracy does not improve. In this static example, the point solution could be averaged to form an improved estimate that would be at least as accurate as the filter solution.

This example shows good performance of a Kalman filter for GPS navigation solutions because it employs accurate models for both the observations and the observer dynamics. In the next two examples, we look at what happens if these rules are violated.

B. Low Dynamics

The next example is for a user traveling along the Earth surface at 100 m/s in the direction due North from 0° latitude, 0° longitude. The position and velocity components of the state are implemented in WGS-84 coordinates, with the full state given by, $\boldsymbol{x}_L \equiv [x\ y\ z \mid \dot{x}\ \dot{y}\ \dot{z} \mid b\ f]^T$.

The process model given in Eq. (17) is used in the filter to propagate the state and covariance between measurement epochs. The discrete process noise matrix for a 1-s measurement interval is assumed to be diagonal with uncertainty for the velocity states set to $(1\ \text{cm/s})^2$, and $(10\ \text{cm})^2$ for the position states.

The initial estimate of the state is

$$\hat{\boldsymbol{x}} = [6{,}377{,}000.0\ \text{m}\quad 3{,}000.0\ \text{m}\quad 4{,}000.0\ \text{m}\quad 0.0\ \text{m/s}\quad 0.0\ \text{m/s}\quad 0.0\ \text{m/s}\quad 0.0\ \text{m}\quad 0.0\ \text{m/s}]^T$$

The diagonal elements of the initial covariance matrix for the position states are each set to 25 km^2; for the velocity states to 1 km^2/s^2; for the clock bias state to 10^4 km^2; and for the clock drift state to $(c \cdot 10^{-6})^2$ km^2/s^2. The receiver takes measurements to all GPS satellites above an elevation of 10 deg at intervals of 1 s. Each pseudorange measurement is assumed to have a variance $R = 36\ \text{m}^2$.

Figure 7 shows the filter estimation error for each of the position and clock bias states. The results are similar to the buoy case in that the filter estimates the trajectory accurately, and the $\pm 1 - \sigma$ uncertainty bounds it computes are reasonable. (In all graphs, the solid line with the "+" symbols represents the actual error. The dotted lines with the "o" symbols represent the filter computed $1 - \sigma$ bounds both before and after each measurement update.)

C. Unmodeled Dynamics

The next step is to introduce unmodeled dynamics. Figure 8 shows the trajectory for a vehicle traveling at a speed of 100 m/s in a 5 km radius circle. Thus, the acceleration is quite mild at 2 m/s^2. The same models were used as in the previous example except that the discrete process noise values were increased to $(1\ \text{m/s})^2$ for each of the velocity states.

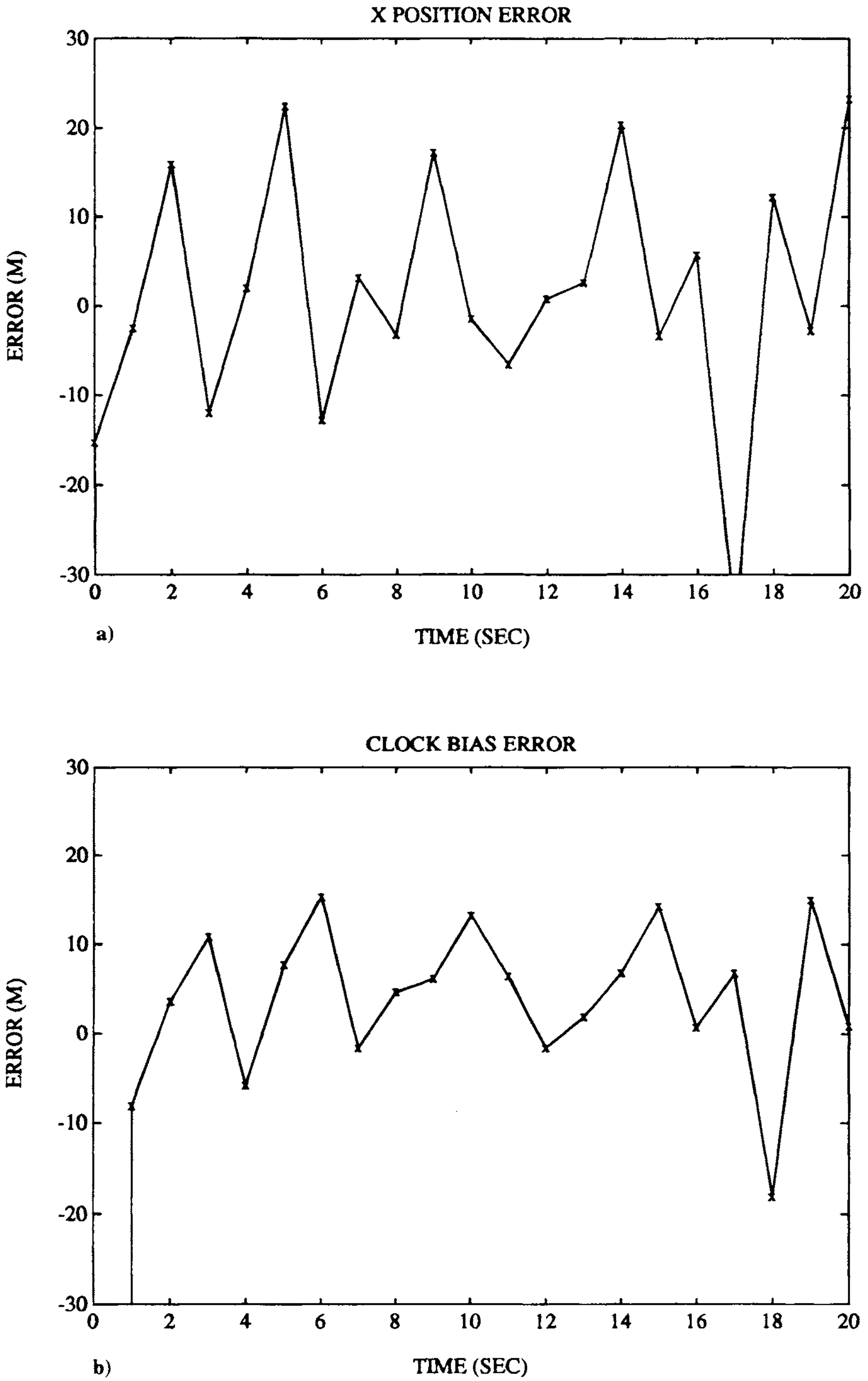


Fig. 6 Point solution estimation errors for buoy example.

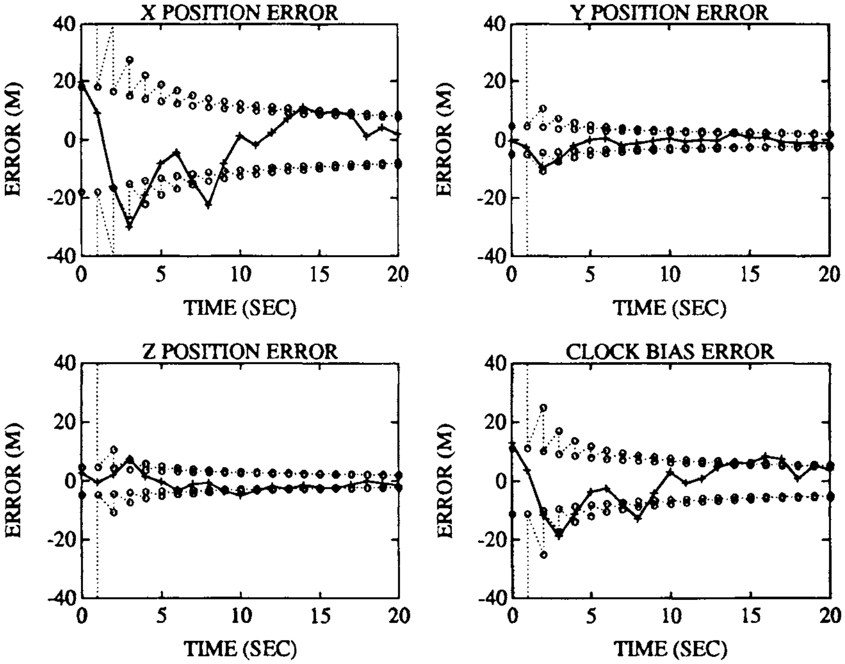


Fig. 7 Kalman filter estimation errors for constant velocity vehicle.

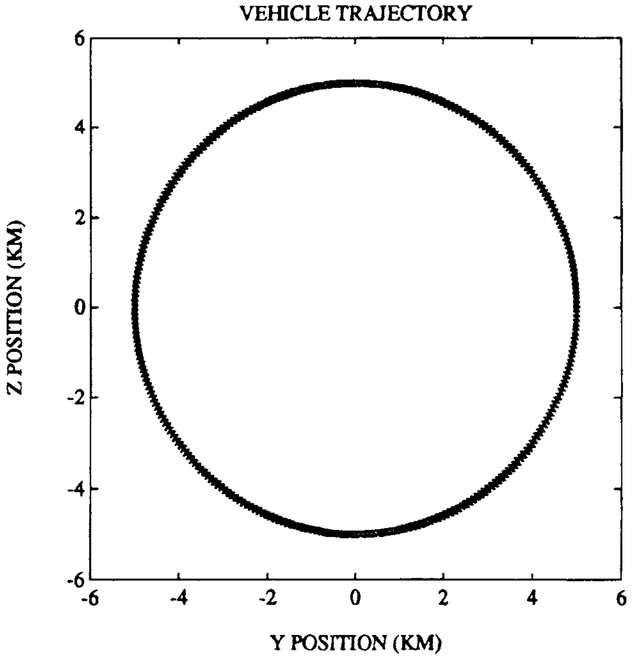


Fig. 8 True *Y–Z* trajectory for vehicle traveling in a circle at a speed of 100 m/s.

The filter results are shown in Fig. 9. Unlike the previous examples, the position and velocity estimation errors for the *Y* and *Z* components have a correlated nature that is not reflected in the filter computed covariance bounds. This illustrates an important limitation of the Kalman filter—it can only work as well as its models. Of course, in this particular case, it would be possible to augment the dynamic model to include acceleration states that would lead to improved performance.

D. Correlated Measurement Errors

As described in Chapter 17 of this volume SA introduces highly correlated errors into the GPS measurements. The figures in Chapter 17 illustrate a variety of observed and simulated SA error profiles. For this example, the model described in Chapter 17, Sec. II was used to generate simulated SA errors for a single satellite. An error profile was created for each GPS satellite, and the buoy filter example was re-run. The only change made to the filter parameters was to increase the measurement variance to $(35\ \text{m})^2$ corresponding to the variance of the actual errors (note that this is not a Schmidt filter). Recall, however, that the filter assumes that the measurement errors are uncorrelated between epochs, which is clearly not the case with SA. Figures 10 and 11 illustrate the filtered and point solution results obtained. Note that in both cases, the solution wanders, following the wandering SA profiles. The filtered solution is smoother, but not more accurate than the point solution.

VII. Summary

This chapter has presented mathematical models for each of the basic GPS measurement types. Pseudoranges were used to form a single point position and clock bias solution. This solution depends only on observations from a single measurement epoch. To improve the navigation accuracy, we must include knowledge of the vehicle dynamics. The extended Kalman filter is an approach commonly used in GPS receivers and data-processing packages. Dynamic models for a stationary, low dynamic, and high dynamic user were given, as well as a useful model for a typical receiver clock. Examples were shown of the performance of the filter as compared to the least squares solution under various conditions, and the consequences of mismodeled dynamics and measurements were described.

In addition to stand-alone navigation, Kalman filtering plays an important role in other GPS applications. Several of the chapters in the companion volume describe models and results obtained for diverse applications such as orbit determination, and aircraft approach and landing. The key to success with this approach is an accurate model of the dynamics and the measurement processes.

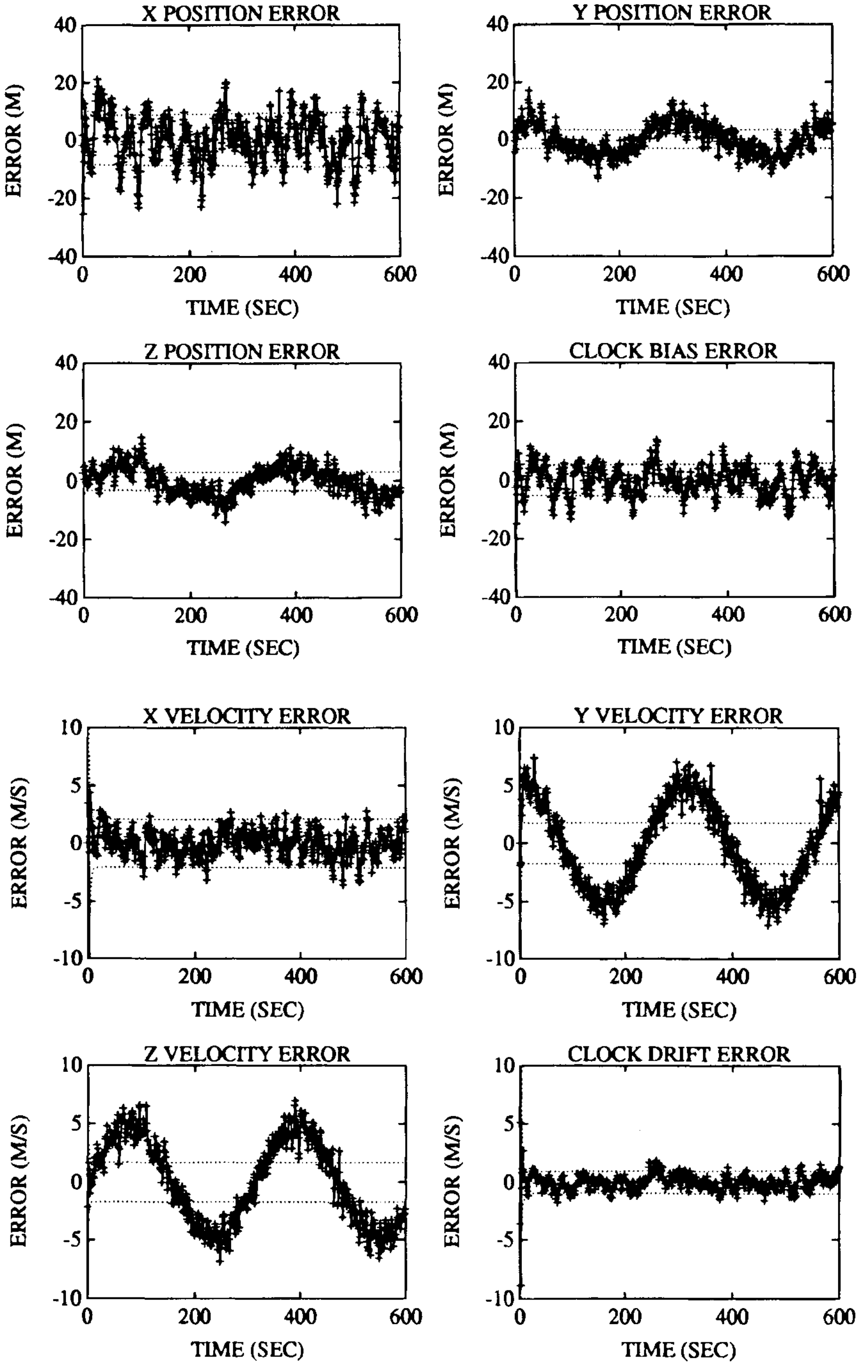


Fig. 9 Kalman filter position, velocity and clock estimation errors for vehicle traveling in a circle. The solid line with the "+" symbols represents the actual error. The dotted lines represent the filter computed 1−σ bounds after each measurement update.

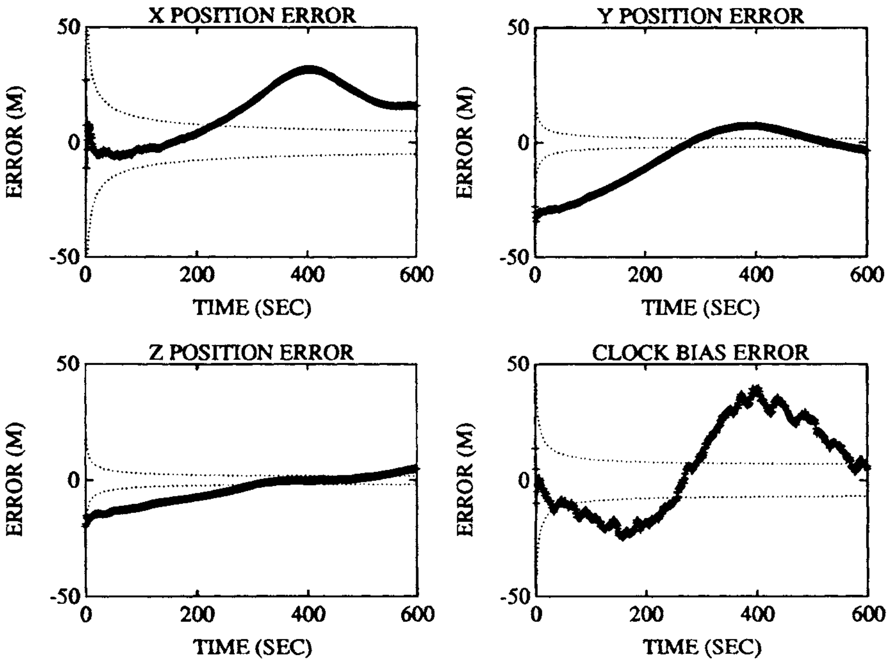


Fig. 10 Kalman filter estimation errors for buoy with SA pseudorange errors. The solid line with the "+" symbols represents the actual error. The dotted lines represent the filter computed 1 − σ bounds after each measurement update. In this case, the filter bounds are clearly not meaningful.

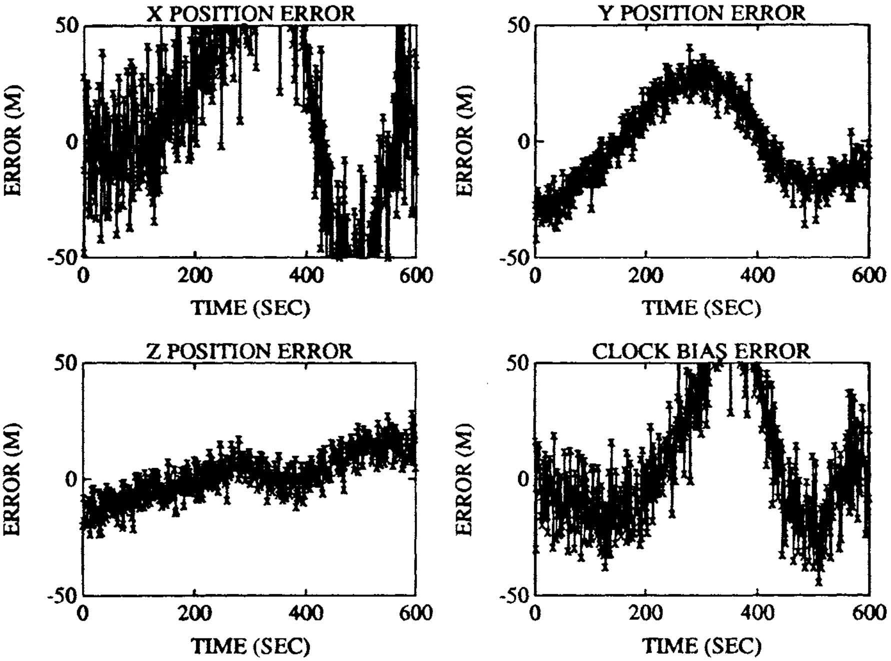


Fig. 11 Point solution estimation errors for buoy with SA pseudorange errors. The solid line with the "+" symbols represents the actual error. The dotted lines represent the filter computed 1 − σ bounds after each measurement update. In this case, the filter bounds are clearly not meaningful.

References

[1]Brown, R. G., and Hwang, P. Y. C., *Introduction to Random Signals and Applied Kalman Filtering,* 2nd ed., Wiley, New York, 1992.

[2]Gelb, A. (ed.), *Applied Optimal Estimation,* MIT Press, Cambridge, MA, 1984.

[3]Bryson, A. E., and Ho, Y. C., *Applied Optimal Control,* Hemisphere, New York, 1975.

[4]Maybeck, P. S., *Stochastic Models, Estimation and Control,* Vol. 1, Academic Press, New York, 1979.

[5]Wells, D. E., et al., *Guide to GPS Positioning,* Canadian GPS Associates, 1986.

[6]Hatch, R., "The Synergism of GPS Code and Carrier Measurements," *Proceedings of the Third International Symposium on Sàtellite Doppler Positioning,* (New Mexico State University), Feb. 8–12, 1982, pp. 1213–1231.

[7]Hwang, P. Y. C., and Brown, R. G., "GPS Navigation: Combining Pseudorange with Continuous Carrier Phase Using a Kalman Filter," *Proceedings of ION GPS-89,* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 27–29, 1989, pp. 185–190.

[8]Anon., ICD-GPS-200, IRN-200B-PR-00J, Rev.B-PR, July 1992.

[9]Golub, G. H., and Van Loan, C. F., *Matrix Computations,* Johns Hopkins University Press, Baltimore, MD, 1989.

[10]Van Dierendonck, A. J., McGraw, J. B., and Brown, R. G., "Relationship Between Allan Variances and Kalman Filter Parameters," *Proceedings of the 16th Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting,* NASA Goddard Space Flight Center, Nov., 1984, pp. 273–293.

[11]Minkler, G., and Minkler, J., *Theory and Application of Kalman Filtering,* Magellan Book Company, 1993.

[12]Kailath, T. (ed.), *Linear Least-Squares Estimation,* Dowden, Hutchinson & Ross, 1977.

[13]Bierman, G. J., *Factorization Methods for Discrete Sequential Estimation,* Academic Press, Orlando, FL, 1977.

Chapter 10

GPS Operational Control Segment

Sherman G. Francisco*
IBM Federal Systems Company, Bethesda, Maryland 20817

THE GPS user solution is based on Space Vehicle (SV) orbit and time scale information contained in the navigation data sets generated by the operational control segment (OCS).[1] The space assets of GPS are managed and supported by this dedicated ground-based segment, which consists of *L*-band facilities to continually track the GPS radio frequency navigation signals, a digital computer network to process scientific and system control data, dedicated *S*-band facilities to conduct duplex information transfer sessions with each individual space vehicle, dedicated secure communications datalinks to couple the globally distributed ground assets into one integrated real-time global system, and a personnel system located at the principal ground facility where dedicated members of the United States Air Force operate the complex system. Functions allocated to this segment of GPS include the operational responsibility to perform the following:

1) Control and maintain the status, health, and configuration of the SV Constellation.

2) Support the user segment with precision predictions of ephemeris and time scale calibration data. Prepare and upload the formatted navigation message data sets to the SV for subsequent metered retransmission to the user community.

3) Monitor the quality of navigation and time transfer services as provided to the end users.

4) Support system interfaces to associated services (i.e., United States Naval Observatory).

5) Manage and schedule the ground assets of the control segment.

This chapter provides a summary description of the present OCS, which includes both the manned master control station (MCS) facility located at Falcon Air Force Base and the automated globally distributed radio frequency station assets deployed to control and continually monitor each satellite of the space vehicle constellation. System architecture, satellite contact coverage, measurement data correction, estimation, and prediction are addressed in this chapter.

*Senior Technical Staff Member.

Figure 1 illustrates the components of this segment and the major system interfaces, including those to other GPS segments and to associated support services.

This OCS architecture provides the GPS operating agency the means for self-sufficient support of the navigation mission once operational space vehicles are stabilized on orbit, and the OCS facilities have ample resources to handle the fully populated constellation supplemented with additional on-orbit spare satellites. External operational interfaces are implemented to the Air Force Satellite Control Facility (AFSCF), to the United States Naval Observatory (USNO), and to the Defense Mapping Agency (DMA) for the respective data exchange functions of SV hand-over in their assigned orbit plane, of coordinating the universal coordinated time (UTC) absolute time scale (which is not visible to GPS instrumentation), and of providing Earth orientation data relative to the existing standard international conventions for inertially fixed coordinate systems. These external support dependencies are strategic in nature because the primary GPS services can autonomously continue service within the established system performance specification for many weeks. Less formal system interfaces also exist to import from the Jet Propulsion Laboratory (JPL) the predicted Sun–Moon position data, to exchange pertinent information with associated user data services, and to support the initial prelaunch activities at Cape Canaveral.

Both the operational procedures and the equipment configuration have been designed to ensure continuous GPS system services. Although the GPS user community is buffered by the distributed store-and-forward data concept implemented in GPS satellite memory, the OCS is designed so that single failures on the ground can be accommodated without disrupting the time-critical functions allocated to this segment. Redundancy of all critical equipment is provided, and

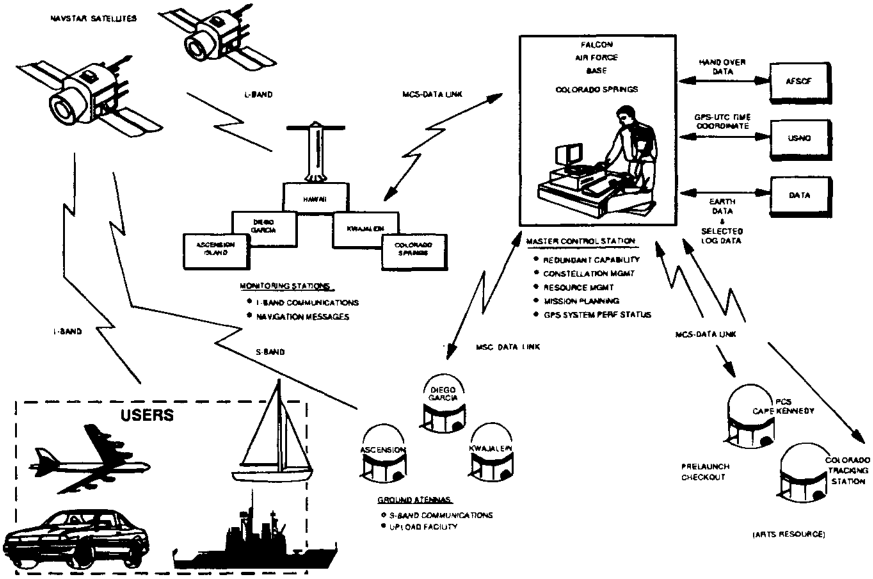


Fig. 1 Operational control segment configuration.

nonreal-time support tasks such as the numerically intensive integration of future reference orbits are identified and judiciously scheduled at OCS convenience on the standby system assets so as to make efficient use of the overall facility investment while meeting the critical data preparation needs. This conservative design philosophy extends to providing database redundancy and to the protection of vital information from contamination. High availability architectural features are cooperatively incorporated in both hardware and software to tolerate single equipment failures, to facilitate equipment reconfiguration and maintenance actions, and to perform (if necessary) process initialization/restart functions without residual corruption of either the database values or of the critical estimation process. Stringent consistency checks are applied to the measurement data, to the estimation process, and to the OCS products generated for internal and for export use. The OCS navigation service product integrity is thus maintained in the presence of normally encountered implementation imperfections.

The robust, underlying GPS service integrity concept of allocating tactical autonomy to each space vehicle and of implementing end-to-end error detection and fail-safe interlocks in the OCS protocols associated with data movement make possible fault-tolerant operational procedures that ensure dataset consistency at the satellite. The occasional OCS process adjustments and equipment reconfigurations encountered during continuous operations are formulated so as to be transparent to users who properly adhere to the total content of the navigation message and the application constraints as documented in the ICD-GPS-200[2] documentation.

Thus, the OCS provides high-integrity navigation data sets at the SV memory. Data errors occurring on the SV–user link are correctable using codes that originated at the ground OCS MCS. Although improbable, an unpredictable flaw in the timing of the GPS signal structure as generated by the SV has occasionally occurred on Block I satellites. The OCS detects and records such events, but has no practical means to provide a service alarm in a timely (10 s) manner. The defensive concept of receiver autonomous integrity monitor (RAIM) has evolved within the community for the user to identify any such SV signal inconsistencies based on the inherent information redundancy of the fully populated constellation. For critical applications such as precision aviation approach navigation, real-time communications can augment the basic GPS service by disseminating both the current status of service and differential corrections to the predictions for each satellite. Such feed-forward error correction overcomes the practical limits of precision time scale and orbit prediction in the baseline GPS concept, but at a cost of service vulnerability.

As space technology continues to advance, some OCS functions of the base system level concept may be reallocated to future space vehicles so as to reduce tactical dependencies on ground assets. Estimation of navigation data (such as SV clock correction term estimation) could be improved by the availability of satellite-to-satellite cross-link measurement capabilities, which makes the onboard refinement of navigation data possible.[3] The precision ranging service from these unaided SVs would also be improved[4] with the potential of a continuously available user range error (URE) of 2 m enabled by the reduced delay in providing precise relative clock corrections to the user sets.

The Consolidated Space Operation Center (CSOC) is located at Falcon Air Force Base (Fig. 2) near Colorado Springs, Colorado. Here the MCS for GPS

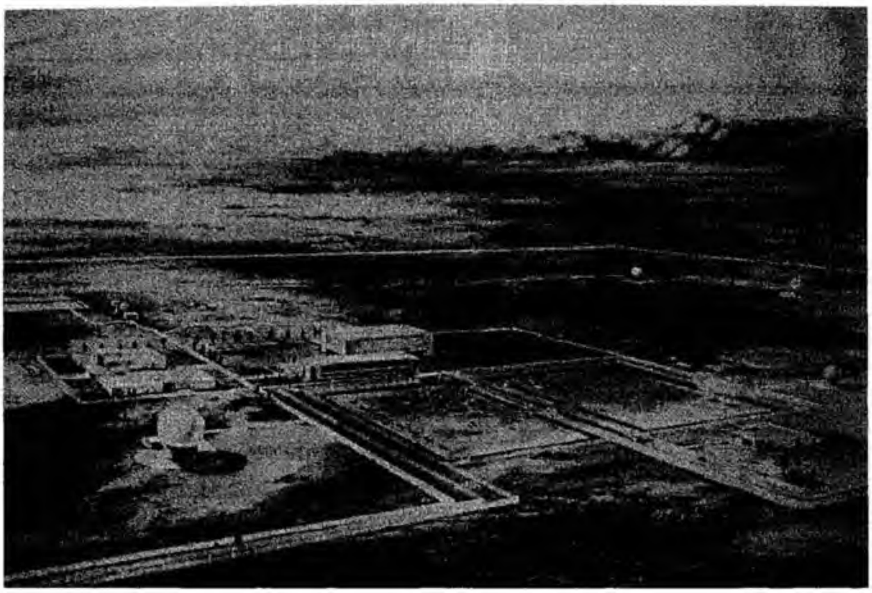


Fig. 2 Consolidated space operations center.

resides and operates 24 hours a day with a dedicated United States Air Force staff. They are responsible for the SV constellation and operate the ground support processing that generates the fresh navigation data uploads.

This facility is the overall operations center for GPS with operational responsibility for all OCS functions including navigation information processing, satellite data upload, vehicle command and control, and overall system management. Full responsibility of each on-orbit SV commences once the normal Earth-oriented attitude stabilization is established which enables *L*-band tracking, an operational capability unique to the GPS control segment (CS). Monitoring the performance of the numerous GPS services and system components by OCS is an essential element of system availability and service quality assurance.

Visibility of the GPS signal structure is provided to the OCS by the five monitor stations (MS). These are remote, unmanned* GPS assets used to passively track the continuous navigation ranging signals of the entire satellite constellation and to acquire the digital data transmitted on *L*-band by each SV. Both the upload message generation and the GPS service quality monitoring are based on analysis of these tracking data.

Another type of unmanned OCS radio frequency asset is the ground antenna (GA), a full-duplex *S*-band communications facility that has dedicated command and control sessions with a single SV at a time. Interactive burst-communication sessions are periodically established for navigation data upload, satellite command, and telemetry reception that are essential to constellation operations and navigation service support. *S*-band ranging services are not available. Both types

*On-call maintenance provided by the host site.

of these distributed GPS radio frequency facilities include the functionality to tolerate automatically short disruption of communications with no induced flaw in service and to support remote reconfiguration with diagnostics conducted from the MCS. Location of these radio frequency facilities has been a programmatic compromise between radio frequency contact utility to the GPS mission and the availability of adequate secure physical facilities with on-call technical support.

I. Monitor Stations

Navigation signal visibility is provided to the MCS by the globally distributed unmanned monitor stations. The GPS monitor stations are radio frequency-passive facilities that receive the same signal structure as the user community. Some low-rate telemetry unique to the instantaneous SV state is of operational interest to the OCS. These stations track the apparent pseudoranges and carrier phase between SV–MS pairs and also collect the transmitted navigation messages for two operationally routine purposes:

1) The pseudorange and carrier phase histories are required to derive precisely and provide the GPS user with ephemerous and clock calibration data for each SV.

2) Both the pseudorange histories and the broadcast navigation data are required to monitor the services as provided to the user community.

Note that the technical performance required of OCS monitor stations must exceed that of the conventional high-precision user set for two fundamental reasons:

1) The OCS task of prediction is more demanding of data quality than is the deterministic end user's current position solution. Both the state separation process encountered in estimating current nonorthogonal system states and the time projection process encountered in predicting future SV states are extremely sensitive to noise, bias, and systematic errors in the measurements. Monitor station measurement errors are magnified in their propagation to become user input error components.

2) The MCS must maintain clear, detailed signal visibility and cope with abnormal signal structures both to permit initial SV process alignments and to support overall system diagnostics. The user has no need to deal with out-of-specification signals or malfunctioning system elements.

The GPS is committed to the operational requirement of global continuous navigation service, and monitoring the system service implies that each satellite's signal should be continuously monitored for pseudorange accuracy compliance, message content, and signal health. The ground projected latitude of block II and later satellites will never exceed the orbit inclination angle of 55 deg; therefore, near-equatorial sites that can track the complete 110-deg groundtrack latitude band are the most desirable MS locations for contact efficiency. Although observations from a spread of latitudes is beneficial to geopotential and force anomaly modeling, covariance error analysis shows that they are not critical to observe the GPS states. Scientific modeling efforts external to the operational program are adequate support to achieve the system performance requirements.

Colorado Springs, Ascension Island, Diego Garcia, Kwajalein, and Hawaii have been chosen as the OCS monitor station sites that best meet the overall OCS requirements, and the precise coordinates of the antenna phase center were accurately determined in the WGS84 coordinate system through special surveys and custom off-line tracking data reductions. Figure 3 illustrates the resulting track coverages provided by this selection of monitor station sites for SV lines of sight that are above the horizontal by a practical 5-deg elevation angle at the receiving antenna. This selection of sites achieves a 95.87 % average coverage for the baseline constellation with only five monitor stations. Contact statistics with each of the operational ground tracks is different, varying from 93.63 to 100% of the ground track. Because of the latitude of Colorado Springs, a slight gap in signal monitoring (not navigation service) does occur for satellites when ground tracks are over the open ocean west of southern South America, which is apparent in Fig. 3. A limited *L*-band capability is also implemented in the eastern launch site at Cape Canaveral for the OCS to verify segment compatibility prior to satellite launch, but use of this facility does not close the slight operational coverage gap.

Regions of tracking coverage overlap (simultaneous *L*-band contact with the same SV by two monitor stations) are very important in establishing a robust GPS estimation process. Observed residuals in pseudorange measurements must be allocated to probable errors in time and in SV position by the action of the Kalman filter estimator. Solution for the states of an isolated satellite and of the monitor stations is quite fragile because of the extensive linear relationships that prevail in pseudorange-based measurement systems and to the effects of accumulated model uncertainty (process noise) when marginal measurement geometry exists to distinguish the error source. Common view strengthens any solution by enabling direct time transfer between MS sites. Fortunately, the extreme situation of sporadic track contact only occurs during system development (and possibly during recovery from an improbable loss of the OCS database). Although it was deemed impractical at the time of initial OCS development to

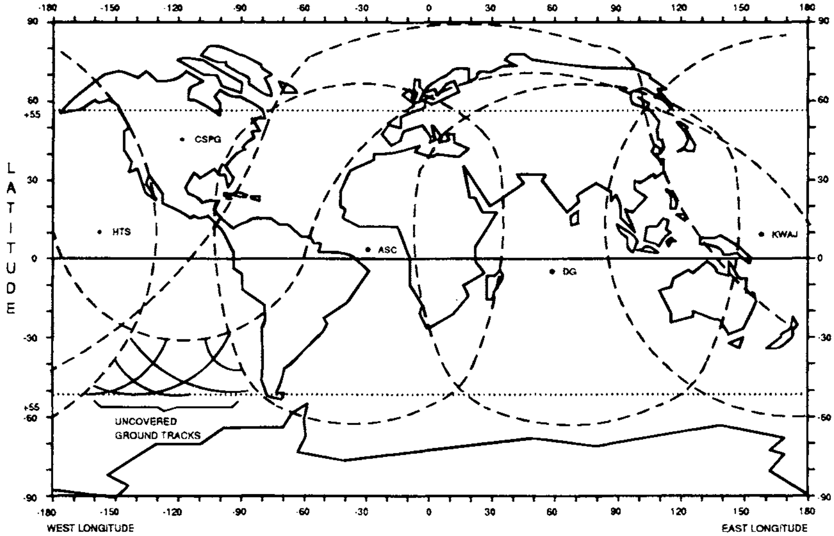


Fig. 3 Monitor station tracking coverage.

implement the total GPS state solution in one large filter formulation so as to fully exploit the available measurement geometry (space and time), near optimum performance has been practically achieved at a much lower computer burden by partitioning the constellation and adhering to the following important system design choices that have been found very beneficial in stabilizing the estimation process and achieve a robust solution within estimation filter partitions, or subsets of states.

1) Select the SVs for inclusion in each partition so as to maintain (near) continuous time transfer between monitor stations within each partition through common-view SV observations. This strategy is made possible by the existence of significant regions of tracking coverage overlap provided by the MS site selection, thus greatly strengthening the GPS system solution by effectively decoupling the MS time and the SV state components of pseudorange errors.

2) Adopt the same SV-signal instance for sampling the common view measurements of a common-view sample set to establish cleanly one satellite state for each common view sample set. This strategy, which effectively decouples satellite velocity states, is achieved in the OCS monitor stations by basing the pseudorange measurement instant on the waveform epoch in the received signal structure. Utilization of this sample time convention reduces the terms in the measurement model, and this choice is essential if corrections for selective availability are to be made with precision.

3) Eliminate the interchannel measurement biases that disrupt the clean separation of system ground states. This reduces the true number of time states and is achieved by the OCS monitor station architecture practice of using, whenever possible, the same physical hardware for common MS functions (i.e., downconverter, station time reference).

Thus, the objective of achieving the full GPS performance potential strongly drove both the selection of MS sites and the detailed receiver equipment architecture. The above design guidance, which greatly influences MS implementation, was formulated through extensive error analysis and simulation efforts at the GPS system level, and these advanced principals have been verified through a decade of stable estimator performance. Even in a partitioned estimator solution, exploiting the common-view information potential is possible through advantageous selection of the MS sites, prudent selection of the SV members within a partition, and the adoption of common-view measurement time tagging (adopting the specific source SV code epoch for sampling). This practice of simultaneous measurement instances for a SV significantly stabilized the partitioned estimator formulation and reduced estimation process sensitivity to systematic error stresses. Errors in position and time are more decoupled, which significantly reduces the degree of error source aliasing. Estimation of the individual satellite trajectories is then numerically solid because the large off-diagonal covariances characteristic of weak geometry are avoided. The relatively large clock process noise levels are less coupled into the position solution, and once the MS time-transfer network closes through common-view measurements, robust estimator performance is established. This strategy has had the effect of immediately stabilizing the relative monitor station time scales for each partition.

Obviously, the precision of knowledge estimated by the CS at the beginning of the prediction period must be an order of magnitude better in quality than the end-ranging service specification after a possibly long prediction interval. Prediction magnifies any errors existing at the beginning of the interval, and so errors in MS measurements affect the user solution more than do the errors in user set measurements. Similarly, the measurements to substantiate system performance metrics must be substantially better than the specified ranging service requirement to be a reliable performance indicator.

Two distinct categories of tracking performance requirements exist. The precision range data, which are used to estimate clock and SV trajectory states, must be free of significant error. Systematic errors are especially damaging because they are not countered by averaging the abundant measurement data, and their presence can cause significant spatial biases in the resulting navigation service product. The OCS allocated processes of separating states and of predicting ahead in time greatly amplifies the size of error to the user. Therefore, systematic errors, such as the effects of multipath interference, must be carefully controlled. The ranging measurement data used by the estimation filters is restricted by the OCS operational procedures to track elevation angles of 15 deg or greater above the horizon so as to avoid significant multipath effects and to eliminate the relatively large, uncorrectable measurement errors that could result from excessively long propagation paths in the troposphere. Exceptional multipath rejection ratio characteristics and phase center stability are still required above this mask angle, but this can be achieved with the simplicity of a single-beam antenna that precludes the problem of antenna channel bias. In contrast, measurement data to support navigation service monitoring and subsequent user process emulation for service verification need only be consistent in quality with that of baseline user practice, which is a relaxation from the accuracy required for the OCS estimation data acquisition. Tracking performance monitoring must, however, extend down to 5-deg elevation angle to achieve coverage with a practical number of MS sites while providing range data better than the baseline user requirement and with acceptable bit error rates.

These OCS *L*-band tracking requirements imposed a formidable antenna design problem. Master station design practice is based on the philosophy that the raw data acquisition performance of an operational system should be consistent, when practical, with the achievable state of the art so as to assure clear visibility of data anomalies for fault analysis. Design studies and then chamber testing of the actual antennas indicate that a phase stability of plus or minus two centimeters is achievable on both L_1 and L_2 without spatial calibration. This uncertainty is consistent with the assessed level of error in the measurement model components.

Thus, a single, unsteered beam fulfilling both the navigation service monitoring and the precision range tracking requirements on both L_1 and L_2 frequencies is implemented at each MS site, and multiple radio frequency channels with resulting equipment duplication and the introduction of interchannel bias problems were avoided. Vega Precision Laboratories provided the high-performance design and developed the production MS antennas (Fig. 4). A shaded, bent turnstile is used to receive dual-frequency, right-hand circularly polarized signal transmission. The conical ground plane with annular chokes at the base produces the specified performance with 14-dB multipath rejection ratio for signal paths above 15-deg elevation angle. The design is optimized for L_2 to compensate for the difference in SV signal transmission strengths. The antenna design contains the low noise

Fig. 4 Monitor station antenna.

amplifier in the pedestal base, as is shown at the Ascension monitor station installation. (An external heated radome with heating is installed at the CSOC location to handle the Colorado winter climate.)

In operational practice, the CS experiences a measurement update innovation level of 2–4 cm for (carrier-aided) 15-min smoothed measurements during normal steady-state operation for each MS. Because this is achieved with estimation filter tuning optimized for the multiday prediction interval performance that demands the use of low but realistic process noise values, the data acquisition and measurement strategy and implementations are verified in practice.

The *L*-band receivers for the CS must have other additional functionality not found in conventional user sets. They must work with signals not in compliance with the operational specification. Initial acquisition of a new satellite cannot be dependent upon the complex signal alignment promised the user, and no correct prompting is assured. The MS receiver must accommodate and detect abnormal signal structures that naturally occur prior to completion of the navigation payload alignment or when a related equipment fault is experienced. Good visibility of the signal structure is retained by the specialized OCS MS receiver design, and the capability for a technical specialist at the MCS to assist in the initial alignment or SV diagnostic process is provided.

For normal tracking, the OCS simultaneously tracks multiple SVs on both *L*-band frequencies to permit correction of the ionospheric delay. Contact scheduling for each MS is based on the requirements for each MS to track all visible satellites above the estimator data criterion of 15-deg elevation angle, and to track continually each SV of the constellation by some MS, if anywhere visible to the OCS. This strategy maximizes the common-view contacts and provides the long track histories for carrier-aided smoothing to support the estimation

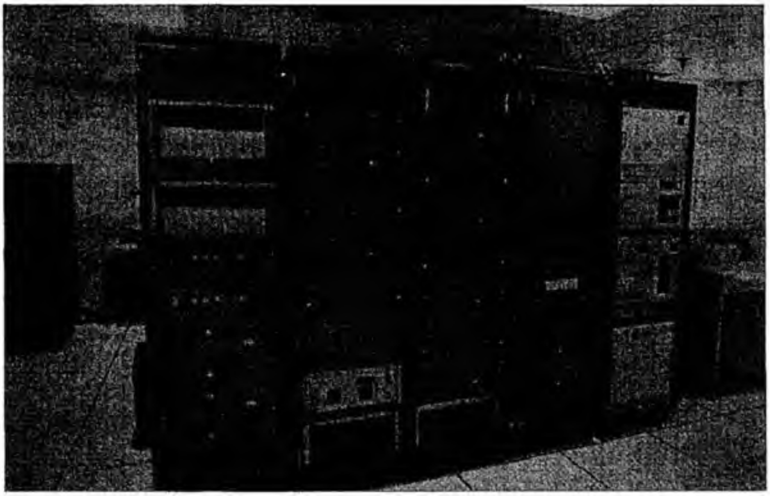


Fig. 5 Monitor station equipment.

process. It also maintains maximum signal monitor contact with each SV. Adequate channel assets are provided in each monitor station to meet this schedule criteria, to provide an active spare channel, and to include a calibration channel programmed to cycle through all the channels. This innovation permits the MCS to sense and correct for any residual interchannel biases in the receiver correlator instrumentation. All modes of track acquisition can be accomplished so that the operator can verify *L*-band support of the numerous user strategies and that reliable SV contact in unusual adverse conditions is possible.

The specialized monitor station receiver equipment was developed by Stanford Telecommunications, Inc. and provide accurate multisatellite range data with minimum interchannel biases. The circa 1978 technology, as shown in Fig. 5, provides data of unprecedented quality to support the precision estimation of clock and ephemeris parameters, and measurement performance is comparable to the best of today's receivers. The internal clock circuits and the two down converters are common to all channels, so that any common bias can be absorbed within the estimated monitor station clock states. Critical cables are trimmed to consistent lengths in order to match delay and drift affects. Pseudorange and accumulated delta range measurements are obtained each 1.5 s on the exact received X-1 epoch for both the L_1 and the L_2 signals.[5] This selection meeting the objective of SV-referenced sampling is convenient to the designer, and the high sample rate captures the information content of the noisy pseudorange measurements. The resulting signal visibility not dependent upon SV alignment is essential to support SV diagnostics and the MS–MCS communications is adequate to return all the data to the MCS. Thermal noise for P-code tracking under the worse specification signal conditions is approximately 2.5 m, one sigma, for pseudorange and less than 0.7 cm, one sigma, for accumulated delta range (ADR). These measurements, the received navigation message, and signal-to-noise power data observations are formatted by the receiver and then transferred to the local digital processor for message handling at this autonomous facility.

A test and calibration signal generator, which provides remote station diagnostic capability, is also controlled by this processor. These test signals can be configured and routed to the LNA, down-converter, or correlator by an MCS ground controller. This capability permits precise station-ready verification and is used to support local maintenance whenever a service call is made.

Frequency Electronic, Inc. (FEI) developed the monitor station frequency standard rack, which establishes the precision reference for pseudorange measurements. Although ultrastable atomic time scales at each MS are not critical to GPS navigation mission once a major portion of the constellation is up, this equipment was important to the development program. Two cesium atomic frequency standards are installed at each unmanned monitor station to permit functional substitution of the hot spare from the MCS. These Hewlett Packard 50-61A Options 4 Standards were modified to provide remote indication of clock parameters considered characteristic of atomic standard health. Phase comparison is instrumented in the FEI design so that the MCS can maintain a current frequency calibration of the backup unit and transfer the timescale to the redundant unit with no further time scale calibration. Remote fault sensing with the option of automatic switchover is provided. The battery-powered backup units have been designed for high availability in unmanned applications. They are fully integrated with the overall MS asset monitoring control structure.

Advanced time subsystems are being introduced to give better support to the time transfer community. This extension is intended to improve the coupling of GPS atomic time and the official USNO reference. Support of local time users can be integrated into the local monitor station reference, which is more current than that available from the SVs.

Remote indicating sensors of barometric pressure, temperature and dew point are included in each MS to permit approximate correction for tropospheric delay. The dew-point sensor at unmanned sites has been undependable, and often a default value is substituted to accommodate sensor malfunction.

Tracking orders and equipment configuration commands are received from the MCS over dedicated, secure communication channels. Measurements and status data are forwarded to the MCS over the same duplex channel, which utilizes commercially developed SDLC protocol to provide error detection and block data retransmission features transparent to the application software. Buffer time of 5 min is provided to accommodate communication equipment reconfiguring or other short service outages without loss of data.

II. Master Control Station

The master control station consists of the processing complex and controller facilities necessary to manage completely the operational GPS space assets and to produce the navigation message. Figure 6 provides an overview of the major functions and the flow of information types throughout OCS. Communications and remote processing can be tested remotely and initialized from the MCS to recover from a fault or extended outage of any remote unmanned facilities. The navigation mission requires upload availability of 98% (not to be confused with

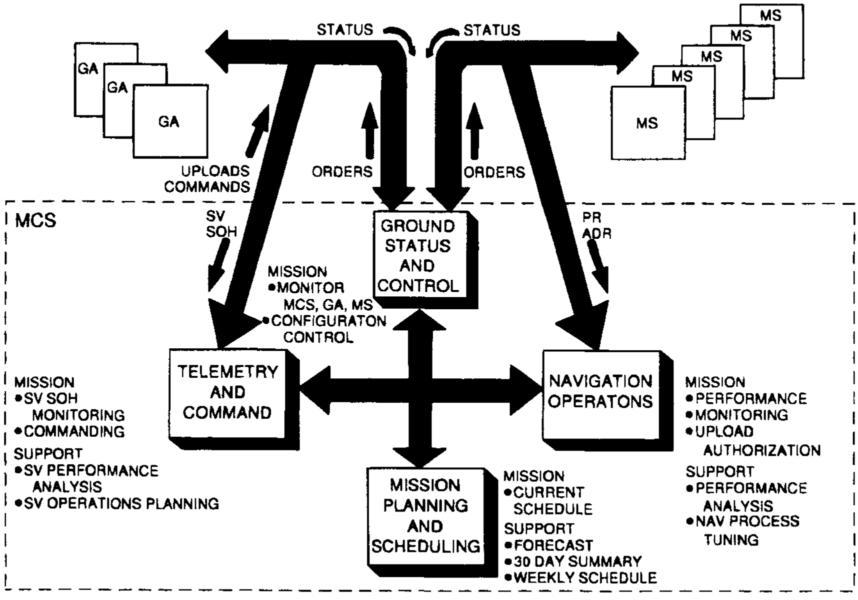


Fig. 6 Operations overview.

navigation service availability), so redundancy is provided for all mission critical equipment. Dual processors, communications controllers, and peripherals are configured to permit processing of the on-line navigation processing and satellite control functions with either processor unit. Many other GPS functions required to maintain the operation are deferrable for short periods of time when necessary to accommodate peak computation loads, fault recovery, and maintenance. Real-time communications are critical, and the OCS has implemented remote initialization and synchronization control from the MCS as well as the diagnostic capabilities to control remotely both digital and analog line looping features. Data Products New England implemented line and service management features of the MCS communications subsystem, which facilitate maintaining communications services. Loop back of the equipment and the control of the communications net is facilitated by the flexibility of this communications interface. All communications with the MCS are encrypted for integrity to prevent spoofing of the system.

Personnel at the MCS control all navigation, processor, constellation, and OCS assets and are responsible for the overall GPS integrity. This requires the establishment of the procedures and efficient access to critical mission data. The MCS provides many identical work consoles initially developed by the Sanders Corporation; each having a keyboard, two-color displays, and a hard copy unit. Air Force controllers are on duty during a full shift. Each shift is under the direction of a senior duty officer, who has overall shift responsibility. This officer is aided by a ground controller, SV engineer, navigation specialist, and multiple pass controllers. The ground controller is responsible for the ground equipment and communications, while the SV engineer is responsible for the overall satellite status and the resolution of problems concerning the SV. The navigation specialist is responsible for the overall navigation mission performance and manages the generation of upload messages. Prime responsibility for the satellite upload transaction, TT&C, and other satellite contacts rests with the pass controllers. Each operator can sign on at any of the consoles; giving his identity, function, and required passwords. The software then configures the display and interface options for that console to meet the specific requirements of the mission function authorized during the sign-on.

The MCS navigation process must generate predicted clock time scale and ephemeris on which to base the data in the navigation message, which is then stored and forwarded to the user by the satellite. Figure 7 illustrates the complex system information flow that is the basis of the GPS concept and integrates the segments to provide the GPS navigation service. As an illustration of the intertwined dependencies within GPS, consider that the pseudorandom radio frequency signal originating at the SV is tracked by the C Segment from which the navigation message is generated and returned to the SV for dissemination to the user, and also back to the CS for verification of the total system performance. The scientific navigation processing that generates the data for the formatted navigation message from measurements acquired by the MS is detailed in a subsequent section of this chapter.

III. Ground Antenna

Navigation message uploads, SV commands, and telemetry data are communicated via *S*-band sessions between a specific SV and one ground antenna (GA) as scheduled by a pass controller. Such information must be communicated to the SV constellation by the network of globally distributed, unmanned ground antenna facilities. Figure 8 illustrates the contact opportunities provided by the chosen GA sites of Ascension Island, Diego Garcia, and Kwajalein. The regions

	INPUT	FUNCTION	PRODUCT
SPACE	NAVIGATION MESSAGE COMMANDS	PROVIDE ATOMIC TIME SCALE GENERATE PSEUDORANGE SIGNALS STORE AND FORWARD NAVIGATION MESSAGES	PSEUDORANDOM RF SIGNAL NAVIGATION MESSAGE TELEMETRY
CONTROL	PSEUDORANDOM RF SIGNAL TELEMETRY UTC	CALIBRATE TIME SCALE, PREDICT EPHEMERIS MANAGE SPACE ASSETS	NAVIGATION MESSAGE COMMANDS
USER	PSEUDORANDOM RF SIGNAL NAVIGATION MESSAGE	SOLVE NAVIGATION EQUATIONS	POSITION VELOCITY TIME

Fig. 7 System information flow.

Next Page

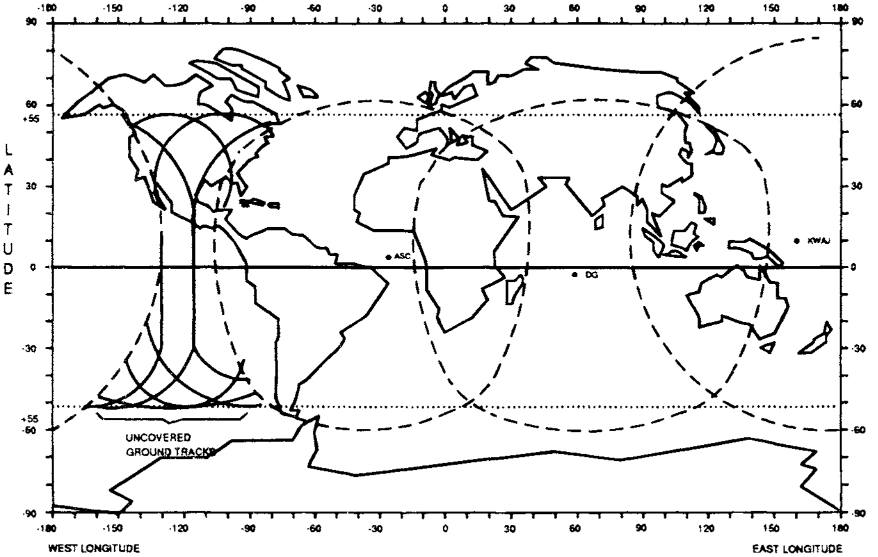


Fig. 8 Ground antenna contact window.

of coverage provide scheduling flexibility to best meet the overall MCS contact requirements. A fourth GA facility owned by GPS is located at the Eastern Launch Site, but is not depicted for coverage because radio frequency transmissions are severely restricted at launch sites, and the prime intent of this equipment is to support segment compatibility verification during prelaunch operations. In the event of a system emergency involving a SV in this contact gap, interoperability with an automated remote tracking station (ARTS) at the Falcon AFB is possible on an emergency basis to provide critical *S*-band services.

The dedicated GA installations developed by the Harris Corporation consist of a Scientific Atlanta 10-m *S*-band antenna and extensive Harris electronic equipment assembled in two rack groups according to function. Figure 9 shows the installation of the rack group containing the radio frequency exciter, high-power transmitter, receiver, and the servodrive equipment. Site selection was influenced by the location of existing remote tracking stations and with an objective of establishing the best overall space-tracking capability. Extensive remote readiness testing, fault alarms, and diagnostic features are incorporated to ensure equipment integrity and to support the on-call site maintenance. On-line equipment safety provisions respond to local sensor inputs to protect the GPS assets from thermal or power anomalies, even if communications with the MCS are disrupted. Command, telemetry, and navigation upload traffic is handled by these GA installations. Tracking orders, equipment configuration commands, and data are received over secured, dedicated duplex communications channels. The local GA processors store this transaction information on disks prior to SV contact, which minimizes operational sensitivity to communications anomalies. The command and upload protocol can be automatically maintained by the GA equipment in accordance with a programmable contact protocol decision tree

Chapter 11

GPS Error Analysis

Bradford W. Parkinson*
Stanford University, Stanford, California 94305

I. Introduction

ALTHOUGH the Global Positioning System (GPS) is clearly the most accurate worldwide navigation system yet developed, it still can exhibit significant errors. By understanding these errors, the user can both hope to reduce them and to understand the limitations of the GPS system. This section of the book should help develop that understanding. This chapter provides both an overview of the sources of error and a detailed analysis of the general error equations. It also presents a standard table of errors that should help clarify the impacts of variations in the specific error magnitudes. Later chapters delve into the expected ranges of these errors.

This development assumes that we are dealing with state-of-the-art receiver technology. In general, this requires a six-channel, continuous tracking implementation. Receivers with fewer channels will probably give a significantly degraded performance. In fact, a number of implementation compromises can produce receiver errors that are greater than those presented here.

This chapter first develops the general error equations and then illustrates how the dilution of precision (DOP) caused by satellite geometry can seriously degrade results. Finally, the error budget is summarized.[1]

II. Fundamental Error Equation

A. Overview of Development

As explained in introductory chapters, a GPS receiver fundamentally measures a quantity called pseudorange ρ, which is a raw, one-way range measurement corrupted by a user clock bias.[2] Using either models or measurements, ρ can be corrected for atmospheric and other effects to produce corrected pseudorange ρ_c. With an approximate user location, the receiver can then process these corrected pseudoranges (to four or more satellites) to determine location in a convenient coordinate system. (Various manufacturers have implemented the "anywhere"

*Professor, Department of Aeronautics and Astronautics; Director of Stanford GPS Program.

fix, which can start from any location. Such techniques are beyond this text and are not a part of error analysis.) This calculation is developed in this chapter.

For GPS, the underlying coordinate system is currently the 1984 World Geodetic System (WGS-84), which is an accepted worldwide geodetic coordinate system. It is expected that this will be replaced with a more accurate version as satellite geodesy improves. It is usual and convenient for the receiver to perform initial calculations in an Earth-centered, Earth-fixed (ECEF) Cartesian coordinate system. These coordinates can then be converted to any other required reference. For *error analysis,* it is usual to consider a local coordinate frame centered at the user and oriented East, North, and Up. This is convenient because many users have differing sensitivities to vertical errors and horizontal errors.

B. Derivation of the Fundamental Error Equation

This section develops the GPS error equation, beginning with the fundamental measurements and proceeding through analysis of the effects of various error sources. It lays the groundwork for the essential understanding of the dilution of precision (DOP) concept.

1. Ideal Measurement

The "true" or ideal measurement is the GPS signal arrival time. This is equal to the signal transmission time delayed by the vacuum transit time and corrected for the true additional delays caused by the ionosphere and troposphere:

$$t_A = t_T + D/c + T + I \tag{1}$$

where t_A = true arrival time (s); t_T = true transmit time (s); D = true range (m); c = vacuum speed of light (m/s); T = true tropospheric delay (s); and I = true ionospheric delay (s).

2. Measured Arrival Time

The measured arrival time reflects the user's clock bias and other measurement errors:

$$t_{Au} = t_A + b_u + v \tag{2}$$

where t_{Au} = arrival time measured by the user (s); b_u = user clock bias estimate (s); and v = receiver noise, multipath, interchannel error (different for each satellite (s)).

3. Satellite Transmission Time

The satellite clock correction transmitted by the satellite can also be in error (the dominant error may be due to selective availability SA):

$$t_{Ts} = t_T + B \tag{3}$$

where t_{Ts} = value of transmission time in the current satellite message (s); and B = true error in satellite's transmission time (includes SA).

4. *True Range*

The true range is the absolute value of the vector difference between the true satellite position and the true user position:

$$D = |\bar{r}_s - \bar{r}_u| = \bar{\mathbf{1}}_s \cdot [\bar{r}_s - \bar{r}_u] \tag{4}$$

where $\bar{r}_s$ = true satellite position; $\bar{r}_u$ = true user position; and $\bar{\mathbf{1}}_s$ = true unit vector from user to satellite.

The right-hand expression in Eq. (4), which uses the vector dot product, is a convenient way to calculate the range in the later user equation. In this calculation, the *estimated* user position can be used to find the unit vector from user to satellite. This unit vector need not be exact. Even *errors* of *several hundred meters* in user or satellite location have a very small effect on the dot product that uses this unit vector. Such a position error would produce angular errors of a few arc-seconds (about 10^{-5} rad) in the unit vector. Because the dot product *error* would be proportional to the cosine of this angle, it is of the order $(\text{angle})^2/2$ or about $10^{-10}/2$ times the range. Because the range is 2×10^7 m, the effect would be less than 1 mm.

5. *Pseudorange*

The user receiver actually "measures" the "pseudorange" ρ, given by the following:

$$\rho = c \cdot (t_{Au} - t_{Ts}) \tag{5}$$

This is called the pseudorange because it is linearly a function of the range to the satellite, but it is also corrupted by the user's clock bias, which must be estimated and removed. In addition, it must be corrected for the estimated satellite time bias and for variations in the speed of transmission.

Substituting Eqs. (3), (2), and (1) into Eq. (5) gives the following result:

$$\rho = D + c \cdot (b_u - B) + c \cdot (T + I + v) \tag{6}$$

Using Eq. (4) in this expression gives the following:

$$\rho = \bar{\mathbf{1}}_s \cdot [\bar{r}_s - \bar{r}_u] + c \cdot (b_u - B) + c \cdot (T + I + v) \tag{7}$$

To account for the estimated value ($\hat{}$) and the estimate error (Δ), each of the above terms is to be broken into two parts as follows:

$\bar{r}_s = \hat{\bar{r}}_s - \Delta\bar{r}_s$, where $\hat{\bar{r}}_s$ = satellite position reported in the transmitted message (m)

$\bar{r}_u = \hat{\bar{r}}_u - \Delta\bar{r}_u$, where $\hat{\bar{r}}_u$ = user estimated position (m)

$\bar{\mathbf{1}}_s = \hat{\bar{\mathbf{1}}}_s - \Delta\bar{\mathbf{1}}_s$, where $\hat{\bar{\mathbf{1}}}_s$ = unit vector from user to the satellite estimated from $\hat{\bar{r}}_s$ and $\hat{\bar{r}}_u$

$b_u = \hat{b}_u - \Delta b_u$, where $\hat{b}_u$ = user clock bias estimate common to a set of simultaneous measurements (s)

$B = \hat{B} - \Delta B - S$, where $\hat{B}$ = satellite transmitted clock bias (s)

ΔB = the "natural" satellite clock error; that is, the error in control system prediction (s)

S = error in transmit time due to SA (s)

$T = \hat{T} - \Delta T$ $\hat{T}$ = estimated (or modeled) tropospheric delay (s)

$I = \hat{I} - \Delta I$ $\hat{I}$ = estimated (or modeled) ionospheric delay (s)

Equation (7) can then be modified to account for the estimated values:

$$\rho_j = (\hat{\bar{\mathbf{1}}}_{sj} - \Delta\bar{\mathbf{1}}_{sj})\cdot(\hat{\bar{r}}_{sj} - \Delta\bar{r}_{sj} - \hat{\bar{r}}_u + \Delta\bar{r}_u) + c\cdot(\hat{b}_u - \Delta b_u - \hat{B}_j + \Delta B_j \\ + S_j) + c\cdot(\hat{I}_j - \Delta I_j + \hat{T}_j - \Delta T_j + \nu_j) \tag{8}$$

where the j subscript is the satellite number and has been added to point out quantities unique to each satellite.

In preparation for conversion to matrix form, Eq. (8) can be rewritten as follows:

$$\underset{(b)}{\underline{\underline{\underline{\hat{\bar{\mathbf{1}}}_{sj}}}}}\bullet\underset{(a)}{\underline{\underline{\hat{\bar{r}}_u - c\cdot\hat{b}_u}}} - \underset{(d)}{\underline{\underline{\underline{\hat{\bar{\mathbf{1}}}_{sj}\bullet\Delta\bar{r}_u + c\cdot\Delta b_u}}}} = \underset{(b)}{\underline{\underline{\underline{\hat{\bar{\mathbf{1}}}_{sj}\bullet\hat{\bar{r}}_{sj} - \rho_j + c\cdot[\hat{I}_j + \hat{T}_j - \hat{B}_j]}}}}$$

$$+ \underset{(c)}{\underline{\underline{\underline{-\hat{\bar{\mathbf{1}}}_{sj}\bullet\Delta\bar{r}_{sj} - \Delta\bar{\mathbf{1}}_{sj}\bullet(\hat{\bar{r}}_{sj} - \hat{\bar{r}}_u) + c\cdot(\Delta B_j + S_j) - c\cdot(\Delta I_j + \Delta T_j) + c\cdot\nu_j}}}}$$

$$+ \text{higher order terms} \tag{9}$$

The large dots represent the dot product. Note that the terms underlined as (a) are the user's position and clock errors to be solved, the terms (b) are estimated or measured by the user, and the terms (c) are unknown errors that produce the solution errors given by (d). The right-hand portion of (b) includes ρ_{cj}, or corrected pseudorange, as: $\rho_{cj} = \rho_j - c \cdot [\hat{I}_j + \hat{T}_j - \hat{B}_j]$. This variable is used below.

Next, define the following matrices for K satellites in view (note that $K = 4$ is normally the minimum number of measurements):

$$\mathbf{G}_{K\times 4} \triangleq \begin{bmatrix} \hat{\bar{\mathbf{1}}}_{s1}^T & 1 \\ \hat{\bar{\mathbf{1}}}_{s2}^T & 1 \\ \vdots & \vdots \\ \hat{\bar{\mathbf{1}}}_{sK}^T & 1 \end{bmatrix} \qquad \mathbf{A}_{K\times 3K} \triangleq \begin{bmatrix} \hat{\bar{\mathbf{1}}}_{s1}^T & & 0 \\ & \hat{\bar{\mathbf{1}}}_{s2}^T & \\ & & \ddots & \\ 0 & & & \hat{\bar{\mathbf{1}}}_{sK}^T \end{bmatrix}$$

$$\hat{\bar{x}}_{4\times 1} \triangleq \begin{bmatrix} \hat{\bar{r}}_u \\ -c\cdot\hat{b}_u \end{bmatrix} \qquad \Delta\bar{x}_{4\times 1} \triangleq \begin{bmatrix} \Delta\bar{r}_u \\ -c\cdot\Delta b_u \end{bmatrix} \qquad \bar{R}_{3K\times 1} \triangleq \begin{bmatrix} \hat{\bar{r}}_{s1} \\ \hat{\bar{r}}_{s2} \\ \vdots \\ \hat{\bar{r}}_{sK} \end{bmatrix}$$

$$\Delta\bar{R}_{3K\times 1} \triangleq \begin{bmatrix} \Delta\bar{r}_{s1} \\ \Delta\bar{r}_{s2} \\ \vdots \\ \Delta\bar{r}_{sK} \end{bmatrix}, \qquad -\hat{\bar{\rho}}_{cK\times 1} \triangleq \begin{bmatrix} -\rho_1 + c\cdot(\hat{I}_1 + \hat{T}_1 - \hat{B}_1) \\ -\rho_2 + c\cdot(\hat{I}_2 + \hat{T}_2 - \hat{B}_2) \\ \vdots \\ -\rho_K + c\cdot(\hat{I}_K + \hat{T}_K - \hat{B}_K) \end{bmatrix} = \begin{bmatrix} -\rho_{c1} \\ -\rho_{c2} \\ \vdots \\ -\rho_{cK} \end{bmatrix}$$

$$\epsilon_{K\times 3K} \triangleq \begin{bmatrix} \Delta\bar{\mathbf{1}}_1^T & & & 0 \\ & \Delta\bar{\mathbf{1}}_2^T & & \\ & & \ddots & \\ 0 & & & \Delta\bar{\mathbf{1}}_K^T \end{bmatrix}, \qquad \bar{P}_{3K\times 1} \triangleq \begin{bmatrix} \bar{r}_u \\ \bar{r}_u \\ \vdots \\ \bar{r}_u \end{bmatrix}$$

and $\Delta\bar{B}$, $\bar{S}$, $\Delta\bar{I}$, $\Delta\bar{T}$, $\bar{v}$ are all obvious. The error in the unit vector to the satellite is $\Delta\bar{\mathbf{1}}_j$. The matrix transpose is a convenient matrix notation for the dot product.

Equation (8) then becomes the following (neglecting higher-order terms):

$$\mathbf{G}\cdot\hat{\bar{x}} - \mathbf{G}\cdot\Delta\bar{x} = \mathbf{A}\cdot\bar{\boldsymbol{R}} - \hat{\bar{\rho}}_c - \mathbf{A}\cdot\Delta\bar{\boldsymbol{R}} + c$$
$$\cdot(\Delta\bar{B} + \bar{S} - \Delta\bar{I} - \Delta\bar{T} + \bar{v}) + \epsilon\cdot(\bar{\boldsymbol{R}} - \bar{P}) \tag{10}$$

The user does not know the last terms of Eq. (10), which are the errors, and calculates position based on the following:

$$\mathbf{G}\cdot\hat{\bar{x}} = \mathbf{A}\cdot\bar{\boldsymbol{R}} - \hat{\bar{\rho}}_c$$

to find the following for $K = 4$:

$$\hat{\bar{x}} = \mathbf{G}^{-1}(\mathbf{A}\cdot\bar{\boldsymbol{R}} - \hat{\bar{\rho}}_c) \tag{11a}$$

and the following for $K > 4$:

$$\hat{\bar{x}} = (\mathbf{G}^T\mathbf{G})^{-1}\boldsymbol{G}^T(\mathbf{A}\cdot\bar{\boldsymbol{R}} - \hat{\bar{\rho}}_c) \tag{11b}$$

using the generalized matrix inverse (or pseudoinverse) of **G**.

These are the fundamental position calculations. Note that **G**, the geometry matrix, is constructed from the set of approximate directions to the satellites, as is the matrix **A**. The vector $\bar{\boldsymbol{R}}$ is constructed from the location of the satellites that has been transmitted and $\hat{\bar{\rho}}_c$ is the corrected pseudorange to each satellite.

Inserting Eq. (11a) back into Eq. (10) cancels appropriate terms and leaves the fundamental error equation*:

$$\mathbf{G}\cdot\Delta\bar{x} = c\cdot(-\Delta\bar{B} - \bar{S} + \Delta\bar{I} + \Delta\bar{T} - \bar{v}) - \epsilon\cdot(\bar{\boldsymbol{R}} - \bar{P}) + \mathbf{A}\cdot\Delta\bar{\boldsymbol{R}} \equiv \Delta\bar{\rho} \tag{12}$$

Thus, the right-hand side consists of all the ranging and calculation errors expressed in meters as we have defined them. Distance and time can be equated by recalling that light travels one meter in about three ns.

[*] The situation for the generalized inverse, Eq. (11b), is somewhat more complicated and is not presented here. The same expression for $\Delta\hat{\bar{x}}$ is obtained. The same generalized development can then be used.

$$\text{for } K = 4, \qquad \Delta\bar{x} = \mathbf{G}^{-1}\Delta\bar{\rho}_c \tag{13a}$$

$$\text{for } K > 4, \qquad \Delta\bar{x} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\Delta\bar{\rho}_c \tag{13b}$$

where $\Delta\bar{x}$ is the positioning error in meters.

III. Geometric Dilution of Precision

It is intuitively obvious that satellite geometry can affect the accuracy of Eq. (13). This section develops the quantitative tools to understand the "dilution" of precision caused by various satellite geometrical configurations.

A. Derivation of the Geometric Dilution of Precision Equation

Now the well-known geometric dilution of precision (GDOP) equation can be easily derived. The covariance of position (m) is calculated as follows:

$$\text{cov(position)} = E(\Delta\bar{x}\cdot\Delta\bar{x}^T) = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\cdot E[(\Delta\bar{\rho}\cdot\Delta\bar{\rho}^T)]\cdot\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1},\ K > 4 \tag{14a}$$

$$\text{cov(position)} = \mathbf{G}^{-1}\cdot E[\Delta\bar{\rho}\cdot\Delta\bar{\rho}^T]\cdot\mathbf{G}^{-T},\ K = 4 \tag{14b}$$

where E is the expectation operator. Because the **G** matrix does not have a random component, it has been brought outside the expectation operation.

If all ranging errors have the same variance $[\sigma_R^2]$ (m^2) and are uncorrelated zero mean ($E[\Delta\rho_i\Delta\rho_j] = 0, i \neq j$), then the expectation in Eqs. (14a and 14b) becomes $\sigma_R^2 \cdot U$, where U is the 4 × 4 identity matrix. Then both Eqs. (14) collapse to the following:

$$\text{cov(position)} = \sigma_R^2\cdot[\mathbf{G}^T\mathbf{G}]^{-1}$$

Therefore, $[\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1}$ is the matrix of multipliers of *ranging* variance to give *position* variance. It is known as the GDOP or geometric dilution of precision matrix. If the position coordinates are the ordered right-hand set, east, north and up, then the *square root* of the ordered diagonal terms from the upper left are: east DOP, north DOP, vertical DOP, and time DOP. Note that σ_R^2 is expressed in m^2 rather than s^2.

cov(Position) =

$$\sigma_R^2(\text{m}^2)\begin{bmatrix} (\text{East DOP})^2 & & \text{covariance} & \text{terms} \\ & (\text{North DOP})^2 & & \\ & & (\text{Vertical DOP})^2 & \\ \text{covariance} & \text{terms} & & (\text{Time DOP})^2 \end{bmatrix}$$

The scalar GDOP is defined to be the square root of the trace of the GDOP matrix. Also HDOP (horizontal) = $\sqrt{(\text{NorthDOP})^2 + (\text{EastDOP})^2}$, and PDOP-(position) = $\sqrt{(\text{HDOP})^2 + (\text{VDOP})^2}$.

For satellites constrained to be above a minimum elevation angle* (greater than 0 deg), the best GDOP for $K = 4$ is obtained when one satellite is overhead and the other three are equally spaced at the minimum elevation angles around the horizon. In later chapters, we explore the power of pseudolites that provide ranging signals *transmitted from the ground.* These have negative elevation angles for an aircraft and can significantly improve geometry.

B. Power of the GDOP Concept

The concept of GDOP is a powerful tool for GPS. All receivers use some algorithm based on GDOP to select the best set of satellites to track among the group of up to 11 satellites in view. *Positioning* accuracy can then be estimated as the *ranging* accuracy multiplied by a dilution factor. This dilution factor (DOP) depends solely on geometry.

Typically, variations in geometry are far greater than variations in ranging accuracy for the nominal satellite constellation. The GDOP concept also quantizes the effect when the nominal satellites are not in view. Examples include local terrain shading, satellite outages, and user shading caused by vehicle extensions such as aircraft wings, etc. During these circumstances, the GDOP calculation for those satellites still being tracked will give the multiplier on ranging accuracies to yield positioning accuracies.

C. Example Calculations

To gain insight into the GDOP concept, some sample calculations are useful. They reveal the tradeoff of accuracy with satellite location. We define satellite direction as azimuth (Az-measured 360 deg clockwise from true north) and elevation (E-measured up from local horizontal—0–90 deg). This can be translated into the east, north, up coordinate frame, and the G matrix becomes:

$$\begin{bmatrix} (\cos(E_1)*\sin(Az_1)) & (\cos(E_1)*\cos(Az_1)) & \sin(E_1) & 1 \\ (\cos(E_2)*\sin(Az_2)) & (\cos(E_2)*\cos(Az_2)) & \sin(E_2) & 1 \\ (\cos(E_3)*\sin(Az_3)) & (\cos(E_3)*\cos(Az_3)) & \sin(E_3) & 1 \\ (\cos(E_4)*\sin(Az_4)) & (\cos(E_4)*\cos(Az_4)) & \sin(E_4) & 1 \end{bmatrix}$$

Using only satellites as ranging sources, the best accuracy† is found with three satellites equally spaced on the horizon, at minimum elevation angle, and one satellite directly overhead (See Table 1).

* Elevation angle is the angle of the satellite above the local horizontal.

† Note that the lower satellite elevation angles tend to have the greater *ranging errors.* The statement is that this configuration offers the best *geometry.* Usually geometry has a larger effect on accuracy than ranging errors.

Table 1 Satellite location

	Sat 1	Sat 2	Sat 3	Sat 4
Elevation, deg	5	5	5	90
Azimuth, deg	0	120	240	0

The following example, for 5-deg minimum satellite elevation, illustrates the technique. Using the best satellite geometry previously described, and the formulas for GDOP we get the following:

$$\mathbf{G} = \begin{bmatrix} 0.000 & 0.996 & 0.087 & 1.000 \\ 0.863 & -0.498 & 0.087 & 1.000 \\ -0.863 & -0.498 & 0.087 & 1.000 \\ 0.000 & 0.000 & 1.000 & 1.000 \end{bmatrix}$$

Recall that the position and time solution ($\mathbf{x}$) is simply given by $\mathbf{G}^{-1}$ (for the four-measurement case*) times the column of corrected pseudoranges ρ_c. $\mathbf{G}$ inverse for the above $\mathbf{G}$ is listed below. Note that each of the first three rows sums to exactly zero. This implies that any common bias in ranging measurements will not affect the position solution. Note also that the last row sums to one. This is the row that gives the time solution (or bias in the local clock). This implies that any common error shows up solely as an error in the local clock. To see these two results, simply multiply on the right by a column whose elements are all the same value. These two results are always true. They arise from inverting the $\mathbf{G}$ matrix, which always has a constant last column of ones.

$$\mathbf{G}^{-1} = \begin{bmatrix} 0.000 & 0.580 & -0.580 & 0.000 \\ 0.670 & -0.335 & -0.335 & 0.000 \\ -0.365 & -0.365 & -0.365 & 1.095 \\ 0.365 & 0.365 & 0.365 & -0.095 \end{bmatrix}$$

The GDOP matrix for this example $(\mathbf{G}^T\mathbf{G})^{-1}$ is shown below. Note that, in this case, all off-diagonal terms are zero except for those correlating vertical errors and time. This correlation is negative, which implies that the errors tend to have opposite signs. In general, large correlations will be found between vertical errors and timing errors, and in general, off-diagonal terms will not be zero.

* For a greater number of satellites, $\mathbf{G}^{-1}$ is replaced with $(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$.

$$(\mathbf{G}^T\mathbf{G})^{-1} = \begin{bmatrix} 0.672 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.672 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.600 & -0.505 \\ 0.000 & 0.000 & -0.505 & 0.409 \end{bmatrix}$$

Listed below are the results of taking the square roots of appropriate diagonal terms of the GDOP matrix above:

HDOP (horizontal DOP) = 1.16 TDOP (time DOP) = 0.64

VDOP (vertical DOP) = 1.26 GDOP = 1.83

PDOP (position DOP) = 1.72

A PDOP of 1.72 is very good. In fact, it is the optimum for four satellites with a minimum elevation angle of 5 deg. A more representative median (50th percentile) worldwide result for PDOP is about 2.5. With all satellites of a 24-satellite constellation available, PDOP numbers as high as six or seven will be found. Thus, geometry can affect the results by a factor of as much as five or more.

D. Impact of Elevation Angle on GDOP

Figure 1 illustrates the impact of satellite position on DOP.[3,4] Maintaining the same symmetric spacing in azimuth, the minimum elevation angle is varied, including negative angles that could only be achieved with Earth-based transmitters (pseudolites).

The HDOP is quite flat over the whole range, which reflects the optimum azimuthal configuration. Satellites more concentrated in one-half of the sky than the other would cause greater variation in these results. Another interesting conclusion is that the lower elevation angles are of significant help for the vertical position.

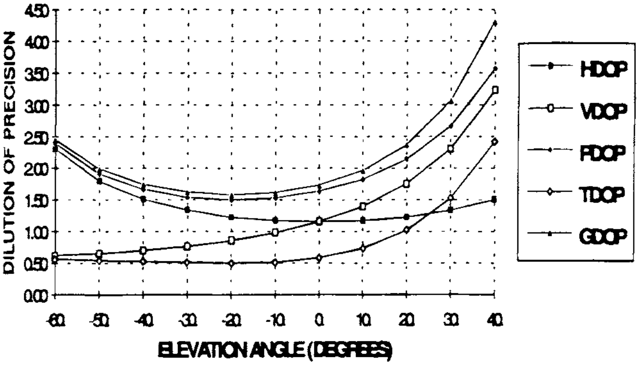


Fig. 1 Dilution of position values for symmetric satellites at various elevation angles above the horizon. The minimum GDOP is at −19.5 deg (i.e., below the local horizontal). This corresponds to locating four transmitters at the apexes of a regular tetrahedron.

IV. Ranging Errors

A. Six Classes of Errors

Ranging errors are grouped into the six following classes:

1) *Ephemeris data*—Errors in the transmitted location of the satellite
2) *Satellite clock*—Errors in the transmitted clock, including SA
3) *Ionosphere*—Errors in the corrections of pseudorange caused by ionospheric effects
4) *Troposphere*—Errors in the corrections of pseudorange caused by tropospheric effects
5) *Multipath*—Errors caused by reflected signals entering the receiver antenna
6) *Receiver*—Errors in the receiver's measurement of range caused by thermal noise, software accuracy, and interchannel biases

Each class is briefly discussed in the following sections. Representative values for these errors are used to construct an error table in a later section of this chapter. A more complete discussion of individual error sources can be found in succeeding chapters.

B. Ephemeris Errors

Ephemeris errors result when the GPS message does not transmit the correct satellite location. It is typical that the radial component of this error is the smallest: the tangential and cross-track errors may be larger by an order of magnitude. Fortunately, the larger components do not affect ranging accuracy to the same degree. This can be seen in the fundamental error Eq. (12). The $\Delta\overline{R}$ represents each satellite position error, but when dot-multiplied by the unit satellite direction vector (in the A matrix), only the projection of satellite positioning error along the line of sight creates a ranging error.

Because satellite errors reflect a position prediction, they tend to grow with time from the last control station upload. It is possible that a portion of the deliberate SA error is added to the ephemeris as well. However, the predictions are long smooth arcs, so all errors in the ephemeris tend to be slowly changing with time. Therefore, their utility in SA is quite limited.

As reported during phase one, (Bowen, 1986) in 1984,[5] *for predictions of up to 24 hours, the rms ranging error attributable to ephemeris was 2.1 m.* These errors were closely correlated with the satellite clock, as we would expect. Note that these errors are the same for both the P- and C/A-codes (see Chapter 16 of this volume for a more detailed discussion of ephemeris and clock errors).

C. Satellite Clock Errors

Fundamental to GPS is the one-way ranging that ultimately depends on satellite clock predictability. These satellite clock errors affect both the C/A- and P-code users in the same way. The error effect can be seen in the fundamental error Eq. (11) as ΔB. This effect is also independent of satellite direction, which is important when the technique of differential corrections is used. All differential stations and users measure an identical satellite clock error.

A major source of apparent clock error is SA, which is varied so as to be unpredictable over periods longer than about 10 minutes. The rms value of SA is typically about 20 m in ranging, but this can change after providing appropriate notice, depending on need. The U.S. Air Force has guaranteed that the two-dimensional rms (2 DRMS) positioning error (approximately 90th percentile) will be kept to less than 100 m. This is now a matter of U.S. federal policy and can only be changed by order of the President of the United States.

More interesting is the underlying accuracy of the system with SA off. The ability to predict clock behavior is a measure of clock quality. GPS uses atomic clocks (cesium and rubidium oscillators),[1] which have stabilities of about 1 part in 10^{13} over a day. If a clock can be predicted to this accuracy, its error in a day ($\sim 10^5$ s) will be about 10^{-8} s or about 3.5 m. The experience reported in 1984 was 4.1 m for 24-hour predictions. *Because the standard deviations of these errors were reported to grow quadratically with time, an average error of 1–2 m for 12-hour updates is the normal expectation.*

D. Ionosphere Errors

Because of free electrons in the ionosphere, GPS signals do not travel at the vacuum speed of light as they transit this region. The modulation on the signal is *delayed* in proportion to the number of free electrons encountered and is also (to first order) proportional to the inverse of the carrier frequency squared ($1/f^2$). The phase of the radio frequency carrier is *advanced* by the same amount because of these effects. Carrier-smoothed receivers should take this into account in the design of their filters. The ionosphere is usually reasonably well-behaved and stable in the temperate zones; near the equator or magnetic poles it can fluctuate considerably. An in-depth discussion of this can be found in Chapter 12, this volume.

All users will correct the raw pseudoranges for the ionospheric delay. The simplest correction will use an internal diurnal model of these delays. The parameters can be updated using information in the GPS communications message (although the accuracy of these updates is not yet clearly established). *The effective accuracy of this modeling is about 2–5 m in ranging for users in the temperate zones.*

A second technique for *dual-frequency P-code* receivers is to measure the signal at both frequencies and directly solve for the delay. The difference between L_1 and L_2 arrival times allows a direct algebraic solution. *This dual-frequency technique should provide 1–2 m of ranging accuracy, due to the ionosphere, for a well-calibrated receiver.*

A third technique is to rely on a near real-time update. An example would be the proposed Wide Area Differential GPS system (WADGPS). *This should also produce corrections with accuracies of 1–2 m or better in the temperate zones of the world.*

E. Troposphere Errors

Another deviation from the vacuum speed of light is caused by the troposphere. Variations in temperature, pressure, and humidity all contribute to variations in the speed of light of radio waves. Both the code and carrier will have the same delays. This is described further in the chapter devoted to these effects, Chapter

13 of this volume. *For most users and circumstances, a simple model should be effectively accurate to about 1 m or better.*

F. Multipath Errors

Multipath is the error caused by reflected signals entering the front end of the receiver and masking the real correlation peak. These effects tend to be more pronounced in a static receiver near large reflecting surfaces, where 15 m or more in ranging error can be found in extreme cases. Monitor or reference stations require special care in siting to avoid unacceptable errors. The first line of defense is to use the combination of antenna cut-off angle and antenna location that minimizes this problem. A second approach is to use so-called "narrow correlator" receivers which tend to minimize the impact of multipath on range tracking accuracies. *With proper siting and antenna selection, the net impact to a moving user should be less than 1 m under most circumstances.* See Chapter 14 of this volume for further discussion of multipath errors.

G. Receiver Errors

Initially most GPS commercial receivers were sequential in that one or two tracking channels shared the burden of locking on to four or more satellites. With modern chip technology, it is common to place three or more tracking channels on a single inexpensive chip. As the size and cost have shrunk, techniques have improved and five- or six-channel receivers are common. Most modern receivers use reconstructed carrier to aid the code tracking loops. This produces a precision of better than 0.3 m. Interchannel bias is minimized with digital sampling and all-digital designs.

The limited precision of the receiver software also contributed to errors in earlier designs, which relied on 8-bit microprocessors. With ranges to the satellites of over 20 million meters, a precision of $1{:}10^{10}$ or better was required. Modern microprocessors now provide such precision along with the co-requisite calculation speeds. *The net result is that the receiver should contribute less than 0.5 ms error in bias and less than 0.2 m in noise.* Further information on receiver errors is available in Chapters 3, 7, 8, and 9 of this volume.

V. Standard Error Tables

These overview discussions on error sources and magnitudes, as well as the effects of satellite geometry, can be summarized with the following error tables. Each error is described as a bias (persistence of minutes or more) and a random effect that is, in effect "white" noise and exhibits little correlation between samples of range. The total error in each category is found as the root sum square (rss) of these two components.

Each *component* of error is assumed to be statistically uncorrelated with all others, so they are combined as an rss as well. The receiver is assumed to filter the measurements so that about 16 samples are effectively averaged reducing the random content by the square root of 16. Of course, averaging cannot improve the bias-type errors.

Finally, each satellite error is assumed to be uncorrelated and of zero mean, so the application of HDOP and VDOP are justified as the last step. Despite these limiting assumptions, the resulting error model has proved to be surprisingly valid. Of course, the assumptions on uncorrelated errors is almost always violated to some degree. For example, if the estimate of zenith ionosphere delay is in error, a proportional error is induced in all measurements through the obliquity calculation. Clearly, such an error would be correlated. These and other correlations have not caused serious problems in the use of this model.

A. Error Table without SA: Normal Operation for C/A Code

Table 2 assumes that SA is not operating. Consequently, the residual satellite clock error, at 2.1 m, is not the dominant error; in fact, the largest error is expected to be the mismodeling of the ionosphere, at 4.0 m. Thus, the worldwide civilian positioning error for GPS is potentially about 10 m (horizontal), as shown in Table 2.

B. Error Table with SA

A second example shows the impact of SA on these errors. Because the deliberately mismodeled clock so dominates the ranging error, all other effects could be safely ignored in the error budget. The results of Table 3 have been repeatedly corroborated by actual measurements. Note that SA is listed as a bias because it cannot be averaged to zero with a 1 s (or less) filter. Selective availability is expected to be zero mean, but only when averaged over many hours or perhaps days. Of course, such averaging is not practical for a dynamic user who only sees the satellite for a portion of the orbit. If differential corrections are used,

Table 2 Standard error model—no SA

	One-sigma error, m		
Error source	Bias	Random	Total
Ephemeris data	2.1	0.0	2.1
Satellite clock	2.0	0.7	2.1
Ionosphere	4.0	0.5	4.0
Troposphere	0.5	0.5	0.7
Multipath	1.0	1.0	1.4
Receiver measurement	0.5	0.2	0.5
User equivalent range error (UERE), rms[a]	5.1	1.4	5.3
Filtered UERE, rms	5.1	0.4	5.1
Vertical one-sigma errors—VDOP= 2.5			**12.8**
Horizontal one-sigma errors—HDOP= 2.0			**10.2**

[a] This is the statistical ranging error (one-sigma) that represents the total of all contributing sources. The dominant error is usually the ionosphere. A horizontal error of 10 m (one-sigma) is the expected performance for the temperate latitudes using civilian (C/A-code) receivers.

Table 3 SA error model

	One-sigma error, m		
Error source	Bias	Random	Total
Ephemeris data	2.1	0.0	2.1
Satellite clock	20.0	0.7	20.0
Ionosphere	4.0	0.5	4.0
Troposphere	0.5	0.5	0.7
Multipath	1.0	1.0	1.4
Receiver measurement	0.5	0.2	0.5
UERE, rms	20.5	1.4	20.6
Filtered UERE, rms	20.5	0.4	20.5
Vertical one-sigma errors—VDOP= 2.5			**51.4**
Horizontal one-sigma errors—HDOP= 2.0			**41.1**

they will eliminate the SA error entirely (if corrections are passed at a sufficiently high data rate) as discussed in Chapter 21, this volume.

The 41-m horizontal error is a one-sigma (σ) result; under the existing agreement between the U.S. Department of Transportation (DOT) and the U.S. Department of Defense (DOD), the 2 DRMS horizontal error is to be less than 100 m. The impact on the vertical error is probably greater, because the VDOP value usually exceeds the HDOP value.

C. Error Table for Precise Positioning Service (PPS Dual-Frequency P/Y Code)

The errors for dual-frequency P/Y code are similar to those above except that SA errors are eliminated because the authorized user can decode the magnitude as part of a classified message. An expected horizontal error is less than 10 m. The ionosphere error is reduced to 1-m bias and about 0.7 m of noise by the dual-frequency measurement. The dominant sources are the satellite ephemeris and clocks. This is illustrated in Table 4.

VI. Summary

Excluding the deliberate degradation of SA, the dominant error source for satellite ranging with single frequency receivers is usually the ionosphere. It is on the order of four meters, depending on the quality of the single-frequency model. For dual-frequency (P/Y-code) receivers (which eliminate SA) the Standard Error Model of Table 1 has one principal change (in addition to the elimination of the SA error). The ionospheric error is reduced from four meters to about one meter.

Greater variations in the errors are due to geometry, which are quantified as dilutions of precision or DOPs. While geometric dilutions of 2.5 are about the worldwide average, this factor can range up to 10 or more with poor satellite geometry. Reduced satellite availability (and consequent increases in DOP) could

Table 4 Precise error model, dual-frequency, P/Y code

	One-sigma error, m		
Error source	Bias	Random	Total
Ephemeris data	2.1	0.0	2.1
Satellite clock	2.0	0.7	2.1
Ionosphere	1.0	0.7	1.2
Troposphere	0.5	0.5	0.7
Multipath	1.0	1.0	1.4
Receiver measurement	0.5	0.2	0.5
UERE, rms	3.3	1.5	3.6
Filtered UERE, rms	3.3	0.4	3.3
Vertical one-sigma errors—VDOP= 2.5			**8.3**
Horizontal one-sigma errors—HDOP= 2.0			**6.6**

be caused by satellite outages, local terrain masking, or user antenna tilting (for example due to aircraft banking). Typical normal accuracy (one-sigma) for well-designed civil equipment under nominal operating conditions *with SA off* should be about 10 m horizontal and 13 m vertical.

References

[1]Martin, E. H., "GPS User Equipment Error Models," *Global Positioning System Papers,* Vol. I, Institute of Navigation, Washington, DC, 1980, pp. 109–118.

[2]Milliken, R. J., and Zollar, C. J., "Principle of Operation of NAVSTAR and System Characteristics," *Global Positioning System Papers,* Vol. I, Institute of Navigation, Washington, DC, 1980, pp. 3–14.

[3]Copps, E. M., "An Aspect of the Role of the Clock in a GPS Receiver," *Global Positioning System Papers,* Vol. III, Institute of Navigation, Washington, DC, 1986.

[4]Massat, P., and Rudnick, K., "Geometric Formulas for Dilution of Precision Calculations," *Navigation,* Vol. 37, No. 4, 1990–1991.

[5]Bowen, R., et al., "GPS Control System Accuracies," *Global Positioning System Papers,* Vol. III, Institute of Navigation, Washington, DC, 1986, pp. 241–247.

Chapter 12

Ionospheric Effects on GPS

J. A. Klobuchar*
Hanscom Air Force Base, Massachusetts 01731

I. Introduction

THE ionosphere is an important source of range and range-rate errors for users of the global positioning system (GPS) satellites who require high-accuracy measurements. At times, the range errors of the troposphere and the ionosphere can be comparable, but the variability of the Earth's ionosphere is much larger than that of the troposphere, and it is more difficult to model. The ionospheric range error can vary from only a few meters, to many tens of meters at the zenith, whereas the tropospheric range error at the zenith is generally between two to three meters. Fortunately, the ionosphere is a dispersive medium; that is, the refractive index is a function of the operating frequency, and two-frequency GPS users can take advantage of this property of the ionosphere to measure and correct for the first-order ionospheric range and range-rate effects directly. Unlike the troposphere, the ionosphere can change rapidly in absolute value. Although the range error of the troposphere generally does not change by more than ±10%, even over long periods of time, the ionosphere frequently changes by at least one order of magnitude during the course of each day. The major effects the ionosphere can have on GPS are the following: 1) group delay of the signal modulation, or absolute range error; 2) carrier phase advance, or relative range error; 3) Doppler shift, or range-rate errors; 4) Faraday rotation of linearly polarized signals; 5) refraction or bending of the radio wave; 6) distortion of pulse waveforms; 7) signal amplitude fading or amplitude scintillation; and 8) phase scintillations.

In order to understand the reasons for these potential effects on GPS performance, a brief description of the major characteristics of the ionosphere is necessary.

II. Characteristics of the Ionosphere

To first order, the ionosphere is formed by the ultraviolet (uv) ionizing radiation from the Sun. Different regions of the ionosphere are produced by different

This paper is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

*Geophysics Directorate, Ionospheric Effects Division, Phillips Laboratory.

chemical species. The ionosphere is a weakly ionized plasma, or gas, which can affect radiowave propagation in various ways. The electron and ion densities are assumed to be equal in the ionosphere, and the density of the ions is much less than 1% of the neutral density at all heights. When the ioriosphere was first discovered, the original regions were named with the alphabetic letters E and F, for electric and field,[1] with the thought that regions of less density, and earlier letters of the alphabet, would eventually be found at lower heights. Today we know that the ionosphere is composed of the D, E, F1, and F2 regions, named in order of increasing height. Figure 1 illustrates the different regions and their electron densities in the ionosphere. The D, E, and F1 regions are closely tied to the uv ionizing daytime radiation from the sun, and are not present at night. The F2 region is present at night, but it is lower in density and generally has its maximum density at a greater height during the night, as compared with daytime.

The various regions of the ionosphere are produced by different wavelengths of radiation from the sun, with the harder solar radiation, namely x rays, penetrating farther into the neutral atmosphere, and the less intense uv radiation being stopped at greater heights where they produce ionization. Much early work was done by Chapman[2] in developing the mathematics of the production of ionization of the atmosphere.

Because the neutral atmosphere is approximately in diffusive equilibrium, the scale height of each neutral atomic and modecular species falls off exponentially with increasing height above the Earth's surface. Thus, the total neutral density is mostly composed of the heavier, molecular species at lower heights, and of the lighter, atomic species at greater heights. Above approximately 180 km, electron diffusion becomes important, and electrons generated by solar uv emissions are free to move to greater heights following the Earth's magnetic lines of force. Additional changes in electron density above approximately 180 km are produced by electric fields that cause electrons to move in a direction perpendicular to the magnetic lines of force, while neutral winds can cause electrons to

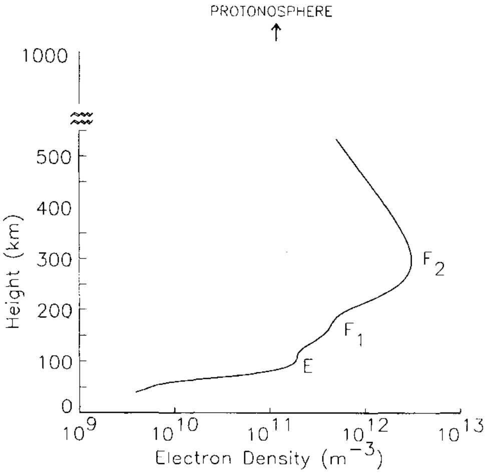


Fig. 1 Electron density of the different regions of the ionosphere vs height for daytime conditions.

flow either up or down the Earth's magnetic field lines, further complicating the specification of electron density at any given height above where diffusion becomes important.

The chemistry of the F2 region is predominately attributable to the ionization of atomic oxygen, whereas the electrons at great heights are attributable to ionized hydrogen gas. The scale height of each species is: $H = kt/mg$, where k is Boltzmann's constant, T is the absolute temperature, m is the mass of the species, and g is the acceleration of gravity.

The scale height of each species is inversely proportional to its atomic weight; thus, the scale height of the electrons associated with H^+ is 16 times greater than that of the electrons due to O^+ ions. Charge neutrality generally is assumed to be the case in the ionosphere, thus the number of electrons always is equal to the number of ions. The scale height of the F2 region is typically 60 km, while that of the protonosphere is over 1,000 km. Thus, although the electron density of the protonosphere is small, the number of electrons does not fall off very fast with increasing height. Therefore, the number of electrons in the protonosphere can be an important fraction of the total, especially during the nighttime periods when the electron density of the F2 region is small, as it normally is during the nighttime.

The parameter of the ionosphere that produces most of the effects on GPS signals is the total number of electrons in the ionosphere. This integrated number of electrons, commonly called the total electron content (TEC), is expressed as the number of electrons in a vertical column having a 1-m^2 cross section, and extending all the way from the GPS satellite to the observer. Details of the behavior of the TEC of the Earth's ionosphere are given in Sec. V of this chapter.

The electron density of the F2 region is not only the highest of the various regions, producing the greatest potential effects on many radiowave systems, but is the most highly variable, safely keeping many ionospheric researchers employed to understand better and be able to predict the physics of its detailed behavior. The major characteristics and importance of each region of the ionosphere for potential effects on GPS signals are summarized as follows (note that heights given are only approximate):

1) D region, 50–90 km: This region, produced by ionization of several molecular species from hard x rays and solar Lyman α radiation, causes absorption of radio signals at frequencies up to the low vhf band, and has no measurable effect on GPS frequencies.

2) E Region, 90–140 km: The normal E region, produced by solar soft x rays, has a minimal effect on GPS. An intense E region, with irregular structure, produced by solar particle precipitation in the auroral region, might cause minor scintillation effects. Sporadic E, still of unknown origin, is very thin and also has a negligible effect at GPS frequencies.

3) F1, 140–210 km: The normal F1 region, combined with the E region, can account for up to 10% of the ionospheric time delay encountered by GPS. Diffusion is not important at F1 region heights, and, as with the normal E region, it has a highly predictable density from known solar emissions. The F1 region is produced through ionization of molecular species, and its electron density nicely merges into the bottomside of the F2 region.

4) F2, 210–1,000 km: The F2 region is the most dense and also has the highest variability, causing most of the potential effects on GPS receiving systems. The height of the peak of the electron density of the F2 region generally varies from 250 to 400 km, but it can be even much higher or somewhat lower under extreme conditions. The F2 region is produced mainly from ionization of atomic oxygen, which is the principal constituent of the neutral atmosphere at those heights. The F2, and to some extent the F1, regions, cause most of the problems for radiowave propagation at GPS frequencies.

5) $H^+ > 1{,}000$ km: The protonosphere, is a region of ionized hydrogen, with a lesser contribution from helium gas. It is of low density, but extends out to approximately the orbital height of GPS satellites. It can be a significant source of unknown electron density and consequent variability of time delay for GPS users. Estimates of the contribution of the protonosphere vary from 10% of the total ionospheric time delay during daytime hours, when the electron density of the F2 region is highest, to approximately 50% during the nighttime, when the F2 region density is low. The electron content of the protonosphere does not change by a large amount during the day, but is depleted during major magnetic storms and can take several days to recover to prestorm values.

III. Refractive Index of the Ionosphere

In order to quantify the propagation effects on a radio wave traveling through the ionosphere, the refractive index of the medium must be specified. The refractive index of the ionosphere, n, has been derived by Appleton and Hartree,[1] and it can be expressed as

$$n^2 = 1 - \frac{X}{1 - iZ - \dfrac{Y^2{}_T}{2(1 - X - iZ)} \pm \left[\dfrac{Y^4{}_T}{4(1 - X - iZ)^2} + Y^2{}_L\right]^{1/2}} \tag{1}$$

where $X = Ne^2/\epsilon_o m\omega^2 = f_n^2/f^2$, $Y_L = eB_L/m\omega, = f_H \cos\theta/f$, $Y_T = eB_T/m\omega = f_H \sin\theta/f$, $Z = v/\omega$, $w = 2\pi f$, where f is the system operating frequency, in Hz and

e = electron charge, -1.602×10^{-19} coulomb
ϵ = permittivity of free space, = 8.854×10^{-12} farad/m
m = rest mass of a electron, = 9.107×10^{-31} kg
θ = the angle of the ray with respect to the Earth's magnetic field
v = the electron-neutral collision frequency
f_H = the electron gyro frequency

The electron gyro frequency f_H is typically 1.5 MHz; the plasma frequency f_N rarely exceeds 20 MHz; and the collision frequency, f_v, is approximately 10^4 Hz. Thus, to an accuracy of better than 1%, the refractive index of the ionospheric is given by the following:

$$n = 1 - (X/2) \tag{2}$$

The ionospheric refractive index is the basis for the effects on GPS signals described later, and the first-order form is sufficient for most purposes. Higher-order corrections are described in Sec. IV.D.

IV. Major Effects on Global Positioning Systems Caused by the Ionosphere

Knowing the refractive index of the ionosphere it is possible to derive the group delay or absolute range error; the carrier phase advance or relative range error; and the Doppler shift or range-rate error. It is also possible to calculate the potential effects of Faraday rotation and refraction, or bending, of the radio wave. The distortion of pulse waveforms is described briefly. Finally, the effects of signal fading, or amplitude scintillation, and phase scintillations are described. All of these effects are produced because the refractive index of the ionosphere differs from unity.

A. Ionospheric Group Delay—Absolute Range Error

1. Single-Frequency Group Delay

The group delay of the ionosphere produces range errors, which can be expressed either in units of distance, or in units of time delay, to GPS users. This group delay can be determined by

$$\Delta t = \frac{1}{c}\int (1 - n)\mathrm{d}l \tag{3}$$

or

$$\Delta r = \int (1 - n)\mathrm{d}l \tag{4}$$

At L-band the first-order refractive index is, $n = 1 - X/2$, where

$$X = \frac{40.3}{f^2} \int N\mathrm{d}l$$

and the ionospheric group delay is

$$\Delta t = \frac{40.3}{cf^2}\int N\mathrm{d}l, \ldots, \text{ seconds} \tag{5}$$

The quantity $\int N\mathrm{d}l$ is the TEC, in el/m^2, integrated along the path from observer to each GPS satellite. The temporal and spatial variations of TEC, which are responsible for the variability of ionospheric time delay to GPS users, are described in Sec. V.

2. Dual-Frequency Group Delay

By measuring the group path delay independently at the two, widely spaced GPS frequencies, $L_2 = f_2$ and $L_1 = f_1$, the TEC along the path from satellite to receiver can be measured directly. A dual-frequency GPS receiver measures the difference in ionospheric time delay at $L_2 - L_1$, referred to as $\delta(\Delta t)$. From Eq. (5), we obtain

$$\delta(\Delta t) = (40.3/c) \times \mathrm{TEC} \times [(1/f_2^2) - (1/f_1^2)] = \Delta t_1[(f_1^2 - f_2^2)/f_2^2] \tag{6}$$

or

$$\Delta t_1 = [f_2^2/(f_1^2 - f_2^2) \times \delta(\Delta t)] \tag{7}$$

where Δt_1 is the ionospheric time delay at L_1.

The value $\delta(\Delta t)$ is obtained from the difference of the simultaneous measurements of the total range, including ionospheric time delay, at the two frequencies f_1 and f_2, because the geometric distance is, of course, the same at all frequencies. The quantity, is $[f_2^2/(f_2^2 - f_2^2)]$ is called the ionospheric scaling factor. For the GPS pair of frequencies, this factor is 1.546.

If the two GPS frequencies had been chosen to be too closely spaced, the differential ionospheric time delay between them would have been very small and would have been masked by the receiving system noise. A wider frequency separation between L_1 and L_2 would have made the measurement capability of absolute ionospheric range error more precise, but probably would have required two separate transmitting and receiving antennas, and more elaborate transmitter and receiver designs. The frequency separation between L_1 and L_2, giving an ionospheric scaling factor of 1.546, is a reasonable compromise between system hardware design and absolute ionospheric range error/time delay requirements. Although Δt_1 is 1.546 times the difference between the two, relatively noisy, pseudorange measurements this differential time delay can be averaged over many samples, and, hence, can be measured to subnanosecond accuracy. The limitations in measuring absolute differential pseudoranges are mostly caused by multipath and lack of a precise knowledge of the differential pseudorange as transmitted from each GPS satellite.

B. Ionospheric Carrier Phase Advance

1. Single-Frequency Ionospheric Carrier Phase Advance

The carrier phase advance, as compared with the received carrier phase in the absence of an ionosphere, can be expressed as

$$\Delta\phi = \frac{1}{\lambda}\int (1 + n)\,\mathrm{d}l, \ldots, \text{cycles, or wavelengths} \tag{8}$$

or

$$\Delta\phi = \frac{f}{2c}\int X\mathrm{d}l = \frac{40.3}{cf}\int N\mathrm{d}l = \frac{1.34 \times 10^{-7}}{f}\int N\mathrm{d}l, \ldots, \text{cycles} \tag{9}$$

Remember that $v_g v_\phi = c^2$, where v_g, v_ϕ are the group and phase velocities, respectively. Although the carrier phase travels faster than the velocity of light, it carries no information, and thus, communication does not occur faster than the velocity of light.

2. Differential Carrier Phase Advance

As a radio signal traverses the ionosphere, the phase of the carrier of the radio frequency transmission is advanced from its velocity in free space. In practice, the amount of this phase advance cannot be measured readily on a single frequency unless both the transmitter and the receiver have exceptional oscillator stability and the satellite orbital characteristics are extremely well known. Usually two, coherently derived, frequencies are required for this measurement. In the case of the GPS satellites, the L_1 and L_2 transmitted carriers are phase coherent, both being derived from a common 10.23 MHz oscillator. The differential carrier phase shift (δ_ϕ) between the two frequencies then can be measured. That differential measurement is related to TEC by

$$\Delta\delta_\phi = [(1.34 \times 10^{-7})/f_L \times [(m^2 - 1)/m^2]/\mathrm{TEC}, \ldots, \text{cycles} \qquad (10)$$

where $m = f_1/f_2$. The GPS system uses differential carrier phase to correct automatically for range-rate errors in its system. Differential carrier phase also provides a very precise measure of changes in relative TEC during a satellite pass, but, because of the unknown number of differential cycles of phase, absolute TEC values must be obtained from the differential group delay measurement.

3. *Relationship Between Carrier Phase Advance and Group Delay*

The relationship between group delay and carrier phase is simply

$$\Delta\phi = -f\Delta t \qquad (11)$$

or, for every cycle of carrier phase advance, there are

$$1/f, \ldots, \text{seconds} \qquad (12)$$

of time delay. In the case of GPS at L_1, one cycle of carrier phase advance is equivalent to 0.635 ns of group delay. The minus sign in Eq. (11) is meant to indicate that the differential code group delay and the differential carrier phase advance move in opposite directions. We must be careful to note this when using carrier-aided code tracking, if the ionosphere changes significantly during the observation period. It is also possible to measure *relative* ionospheric changes using the L_1 C/A code minus L_1 carrier phase during a pass.

4. *Useful Numbers for GPS Are Shown in Table 1.*

C. Higher-Order Ionospheric Effects

Because we are concerned here only with radiowave propagation at GPS frequencies, the terms X, Y_T, Y_L, and Z in Eq. (1) are all much less than one. Thus, the refractive index of the ionosphere at GPS frequencies, as given by Brunner and Gu,[3] neglecting those terms whose magnitude is less than 10^{-9}, can be expressed as

$$n = 1 - (X/2) \pm (XY/2)\cos\theta - (X^2/8) \qquad (13)$$

As Eq. (13) illustrates, the terms contributing to the refractive index of the ionosphere are A: 1, the free-space velocity; B: $(X/2)$, the first-order, or $(1/f^2)$ term; C: $(XY/2)\cos\theta$, the second-order, or $(1/f^3)$, term; and D: $(X^2/8)$, the third-order, or $(1/f^4)$ term where f is the GPS operating frequency. X and Y are defined in Sec. III. Using the Brunner and Gu[3] derivation, the magnitudes

Table 1 Relationships between the various global positioning system first-order measured parameters and total electron content of Earth's ionosphere

L_2–L_1, differential group delay

360 deg, or 151.098 ns of delay, measured at L_1, or 97.75 ns of differential delay; i.e., 1 code chip

360 deg = 278.83×10^{16} (el/m^2); 1 deg = 0.7745×10^{16} (el/m^2)

1 ns of differential code delay
= 2.852×10^{16} (el/m^2)
= 1.546 ns of delay at L_1
= 0.464 m of range error at L_1

1 ns of delay, measured at L_1
= 1.8476×10^{16} (el/m^2)
= 0.300 m of range error at L_1

1 cycle, or 1 wavelength, 19.04 cm, of carrier phase advance at L_1 = 1.173×10^{16} (el/m^2)

1 m of range error:
measured at L_1 = 6.15×10^{16} (el/m^2)
measured at L_2 = 3.73×10^{16} (el/m^2)

1 TEC unit [1×10^{16} (el/m^2)]
= 0.351 ns of differential delay
= 0.542 ns of delay at L_1
= 0.163 m of range error at L_1
= 0.853 cycles of phase advance at L_1

L_2–L_1 differential carrier phase advance, measured at L_2

1 deg = 6.456×10^{13} (el/m^2)
0.1 rad = 3.699×10^{14} (el/m^2)
360 deg = 2.324×10^{16} (el/m^2)

To convert differential carrier phase advance (measured at L_2) to an equivalent single frequency phase change at a specified frequency.

Frequency	Multiply GPS differential carrier phase by
244 MHz	12.81
1 GHz	3.125

of the higher-order terms at GPS frequencies, for maximum worldwide ionospheric conditions, are $B \approx 2 \times 10^{-4}$; $C \approx 2 \times 10^{-7}$; and $D \approx 2 \times 10^{-8}$.

The ratios of the higher-order terms, C and D, to the first order term, B, again under these worst case ionospheric conditions, is $C/B \approx 10^{-3}$; and $D/B \approx 10^{-4}$.

Thus, these higher-order terms are much less than 1% of the first-order term at GPS frequencies, even for the extremely high value of $f_n = 25$ MHz used for the maximum ionospheric plasma frequency, and for the GPS L_2 frequency used in the computation. For a more typical ionospheric maximum plasma frequency of, say, 12 MHz, and for GPS L_1 users, the higher-order terms in ionospheric refraction are even much less than those given here. Thus, within better than 0.1% accuracy, even during worst case ionospheric conditions, the ionospheric refractive index at GPS frequencies can be expressed simply as

$$n = 1 - (X/2) \tag{14}$$

During times of high TEC, the first-order range error can be a few hundred meters. At these times, higher-order ionospheric effects can be several tens of centimeters of range error, which represent large errors in geodetic measurements. Brunner and Gu[3] have used the full form of the refractive index given in Eq. (1) to calculate the residual range error from the first-order form for refractive index. Their model also includes the geomagnetic field, and the effects of ray bending at both the GPS frequencies, L_1 and L_2. They claim that their improved form of dual-frequency ionospheric correction eliminates the ionospheric higher-order effects to better than 1 mm residual range error. However, in order to achieve this order of ionospheric error correction, they require knowledge of the actual maximum electron density, N_m, an ionospheric electron density profile shape factor they call η, and the average value of the longitudinal component of the Earth's magnetic field along the ray path, $\overline{B(\cos\,\theta)}$. In a practical case, these parameters are not easy to estimate.

Bassiri and Hajj[4] have done similar work on higher-order ionospheric range errors for GPS. They find the magnitudes of the second- and third-order terms at L_1 to be ~ 1.6 cm and ~ 0.9 mm, respectively, for a TEC of 10^{18} (el/m^2). Rather that requiring a knowledge of the electron density profile shape, as is the case with Brunner and Gu,[3] they use a constant ionospheric shape factor of 0.66, and they assume a constant height for the maximum of the electron density profile. Their form of higher-order ionospheric corrections are not as good as those claimed by Brunner and Gu,[3] but they are much easier to implement, and, in the practical case of actual ionospheric data required, may result in corrections of the same magnitude. GPS users who require ionospheric corrections to centimeter or millimeter accuracy should refer to the works of Brunner and Gu[3] and Bassiri and Hajj[4] for additional details.

D. Obtaining Absolute Total Electron Content from Dual-Frequency GPS Measurements

1. Removing Multipath Effects

If both the differential carrier phase and the differential group delay are measured with a dual-frequency GPS receiver, the user easily can obtain both the absolute TEC and its rate of change. Jorgensen,[5] first showed that the differential group delay could be used to fix the differential carrier phase to an absolute scale after a satellite pass, thus obtaining the best of both measurements; namely, the absolute scale obtained from the differential group delay and the precision of the differential carrier phase. Hatch[6] suggested that combining GPS carrier with code measurements results in better absolute and relative positioning. Because the differential carrier phase is much less sensitive to multipath, the final result, providing there are no unresolved carrier cycle slips during the pass, is a precise, smooth rendition of absolute TEC over an entire pass. The ratio of multipath effects on the rf carrier, as compared with multipath observed on the P-code modulation, is proportional to their respective wavelengths. That is, multipath on the L_2 carrier is 120 times smaller than on the 10.23 MHz modulation.

Figure 2 shows a typical pass of a GPS satellite. The top panel is the L_1 carrier signal strength divided by the receiver noise, on a linear amplitude scale. The middle panel shows the relative differential carrier phase on a relative scale.

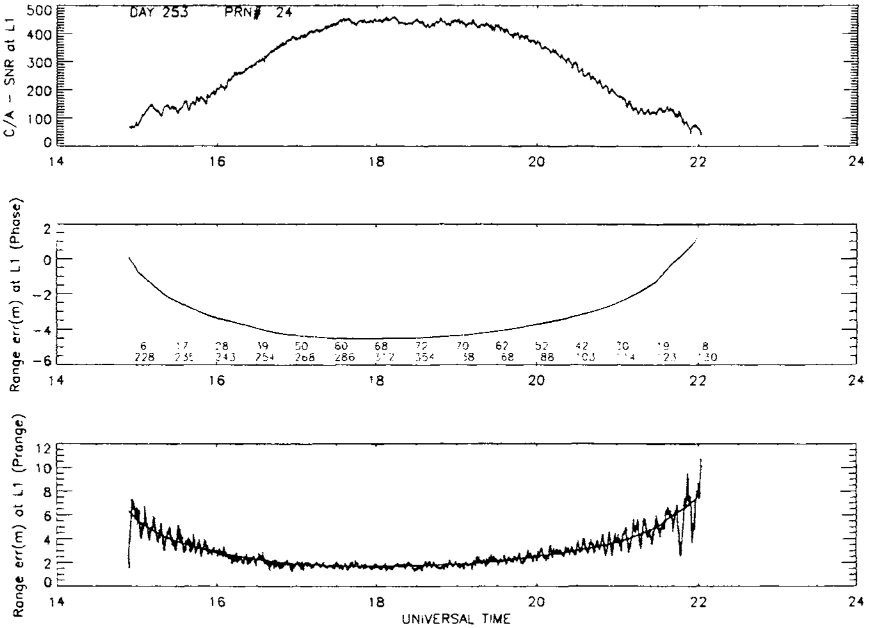


Fig. 2 Recording of a typical global positioning system pass, showing the carrier-assisted relative signal-to-noise ratio at L_1 (top panel), relative range error computed from the differential carrier phase (middle panel), and absolute rate error computed from the differential P-code group delay (bottom panel).

Normally the differential carrier phase is set at zero relative range error at the beginning of each pass. Elevation and azimuth values are printed in the middle panel, just under the relative differential phase record. The bottom panel shows the absolute ionospheric range error, in meters, obtained from differential group delay. Note the large multipath effects in the differential group delay, especially at both ends of the pass. The fact that the differential group variations are attributable to multipath effects becomes obvious if two, or more, successive days of passes are plotted to the same scale. The individual multipath variations agree exactly with a time shift corresponding to the sidereal difference in the time of passage of the GPS satellite along the same track over successive days.

The bottom portion of Fig. 2 also shows the differential carrier phase, now translated to an absolute scale. The process of converting the differential carrier phase to an absolute scale, by fitting to the absolute differential group delay, is done over the higher-elevation portions of each pass, when multipath is generally smallest. Multipath may not be a zero mean process and may not necessarily average out if the fitting procedure were done over the entire pass, or even worse, if the fitting were done only over the low-elevation portions of a pass.

After the differential carrier phase is converted to an absolute scale by fitting it to the differential group delay curve over the desirable, low multipath, portion of each pass, the differential group delay data are simply discarded, because they have done their task. The final TEC values are precise, accurate, and without multipath, unless the multipath environment is really terrible, in which case a small, residual amount of multipath can even be seen in the differential carrier phase.

If a user requires absolute TEC measurements soon after acquisition of each GPS satellite at a low-elevation angle, the fitting of the differential carrier phase to the differential group delay must be done for only a few minutes of data, usually where multipath on the differential group is large. There are several, relatively straightforward, schemes for removing multipath at low-elevation angles. One is to use a high-gain, directional antenna that has smaller multipath effects. An antenna with a large ground plane also can attenuate multipath reflections. Another method is to take advantage of the fact that the orbit of each GPS satellite repeats in its ground track over successive days, thereby giving the stationary user nearly the same multipath conditions. Suitable matching and filtering techniques can be employed to use the multipath for preceeding days to remove its effects in near real time. The user must take care to see that the day-to-day differences in multipath for each GPS satellite really are the same. GPS antennas deployed in locations where the multipath effects may change from day to day, such as near automobile parking lots or aircraft terminals, may not be suitable for this procedure.

2. *Automated Ionospheric Range and Range-Rate Error Corrections*

In using an operational dual-frequency GPS receiver, all the processes described above are done automatically and are transparent to the user. If the ionospheric first-order range and range-rate error corrections are done for each satellite soon after each satellite is acquired, it is likely that only the differential group delay is used to obtain an absolute range error correction. However, the process of determining an absolute ionospheric correction by using both the differential carrier phase and the differential group delay, with an improving fit of the phase data to the group delay data as more data are acquired, could be used routinely.

E. Ionospheric Doppler Shift/Range-Rate Errors

Because frequency is simply the time derivative of phase, an additional contribution to geometric Doppler shift results because of changing TEC. This additional frequency shift is generally small compared with the normal geometric Doppler shift, but can be computed by

$$\Delta f = (\mathrm{d}n/\mathrm{d}t) = [(1.34 \times 10^{-7})/f]\,(\mathrm{d}/\mathrm{d}t)\,\mathrm{TEC}, \ldots, \mathrm{Hz} \qquad (15)$$

For high-orbit satellites, such as the GPS satellites, where the apparent satellite motion across the sky is slow, the diurnal changes in TEC are generally greater than the geometric ones. An upper limit to the rate of change of TEC, for a stationary user, is approximately 0.1×10^{16} (el/m^2) per second. This value yields an additional frequency shift of 0.085 Hz at L_1, which would not be significant compared with a typical required receiver carrier tracking loop bandwidth of at least a few Hz. The value of 0.085 Hz at L_1 corresponds to 1.6 cm/s of range-rate error.

Ionospheric range-rate or Doppler shift errors are attributable to the time rate of change of the electron content of the ionosphere, as seen by the observing system. The range-rate error depends upon the diurnal rate of change of the electron content of the ionosphere, the structure of any large-scale irregularities that may exist in the region, and the motion of any vehicle. For instance, a GPS satellite moving up from the horizon will usually encounter fewer electrons as it rises in elevation, simply because of the decrease of the signal path length in the ionosphere. An observer in a high-velocity aircraft, or even more so, in a low-orbit spacecraft, generally will encounter geometric changes far greater than the temporal rate of change of electron content in the ionosphere.

For the slowly moving GPS satellites, the satellite motion, diurnal changes in the ionosphere, and observer platform motion, all contribute to ionospheric range-rate errors. It is difficult to model the relatively high day-to-day variability of range-rate changes attributed to the ionosphere even for a fixed observer because of the large variability in the day-to-day ionospheric rates of change. Thus, for ionospheric range-rate errors, corrections through the use of an ionospheric model, particularly one with a simple representation of diurnal ionospheric changes, such as that in the single-frequency GPS user ionospheric algorithm, are not recommended. The dual-frequency GPS user can, of course, automatically correct for both the first-order range and range-rate ionospheric errors.

F. Faraday Rotation

1. Amount of Faraday Rotation

When a linearly polarized radio wave traverses the ionosphere, the wave undergoes rotation of the plane of this linear polarization. At frequencies of approximately 100 MHz, and higher, the amount of this polarization rotation can be described by

$$\Omega = \frac{k}{f^2} \int B \cos \theta \, N dl, \ldots, \text{radians} \tag{16}$$

where the quantity inside the integral is the product of electron density times the longitudinal component of the Earth's magnetic field, integrated along the radio wavepath. Many ionospheric workers have used this effect to make measurements of the TEC of the ionosphere.

Because the longitudinal magnetic field intensity, $B \cos \theta$, changes much slower with height than the electron density of the ionosphere, Eq. (16) can be rewritten as

$$\text{TEC} = (\Omega f^2 / k B_L) \tag{17}$$

where $B_L = \overline{B \cos(\theta)}$ is taken at a mean ionospheric height, usually near 400 km, $k = 2.36 \times 10^{-5}$, and TEC is simply $\int N dl$.

Generally, the equivalent vertical TEC is determined by dividing the slant TEC by the secant of the zenith angle at a mean ionospheric height. The equivalent vertical TEC is the one most often used for comparison purposes among sets of TEC data, because these different slant TEC values cannot easily be directly compared. Much of the TEC data available today from stations throughout the world, used in model construction and testing, are from Faraday rotation measurements from vhf telemetry signals of opportunity from various geostationary satellites.

For satellite navigation and communication designers, the Faraday polarization rotation effect is a nuisance. If a linearly polarized wave is transmitted from a satellite to an observer on, or near, the surface of the Earth, the amount of polarization rotation may be nearly an odd integral multiple of 90 deg, thereby giving no signal on the receiver's linearly polarized antenna, unless the user is careful to realign the antenna polarization for maximum receiver signal. The Faraday rotation problem is overcome by the use of circular polarization of the correct sense at both the satellite and at the user's receiver.

2. *Faraday Rotation Effects on Global Positioning Systems*

GPS signals are transmitted with right-hand circular polarization; thus, Faraday rotation is not a problem with GPS users. The optimum receiving antenna for GPS users also would be one of right-hand circular polarization, to ensure that the receiving antenna polarization matches the characteristics of the received signal. However, it is impossible to design a nearly omnidirectional GPS receiving antenna having circular polarization over most of the entire sky. If a GPS user uses a linearly polarized antenna, the loss will be 3dB, because only one-half of the potential signal energy is being received. Thus, the optimum receiving antenna will exhibit right-hand circular polarization over as much of the sky as possible. Generally the mobile user finds it difficult to utilize circular polarization, because of the continual vehicle directional changes; thus, the user settles for nearly linear polarization. The up to 3-dB loss between transmitted circular polarization and receiver nearly linear polarization is a necessary price GPS users pay for antenna maneuverability and simplicity. The transmitted signal levels from the GPS satellites were designed to provide adequate signal strength for users with linearly polarized antennas.

Had the GPS satellites been transmitting linearly polarized radio waves, the polarization rotation, viewing satellites in various directions from a northern midlatitude station, could be up to 90 deg. For values of polarization rotation near 90 deg, signal loss for two, cross-aligned, linearly polarized antennas can be in excess of 30 dB. Thus, circular polarization was wisely chosen for use on GPS.

G. Angular Refraction

The refractive index of the Earth's ionosphere is responsible for the bending of radio waves from a straight line geometric path between satellite and ground. Normally, for GPS users, the small bending of radio waves is not a problem. The angular refraction, or bending, produces an apparent higher elevation angle than the geometric elevation. Millman and Reinsmith[7] have derived expressions relating the refraction to the resultant angular bending. Perhaps the easiest expressions to use, as given by Millman and Reinsmith,[7] relate the ionospheric range error to angular refraction. This expression is:

$$\Delta E = \{(R + r_0 \sin E_0)\,(r_0 \cos E_0)/[h_i(2r_0 + h_i) + (r_0^2 \sin^2 E_0)] \times (\Delta R/R)\} \quad (18)$$

where E_0 is the apparent elevation angle; R is the apparent range; ΔR is computed from, $\Delta R = (40.3/f^2) \times \mathrm{TEC}$; r_0 is the Earth's radius; and h_i is the height of the centroid of the TEC distribution, generally taken to be between 300 and 400 km.

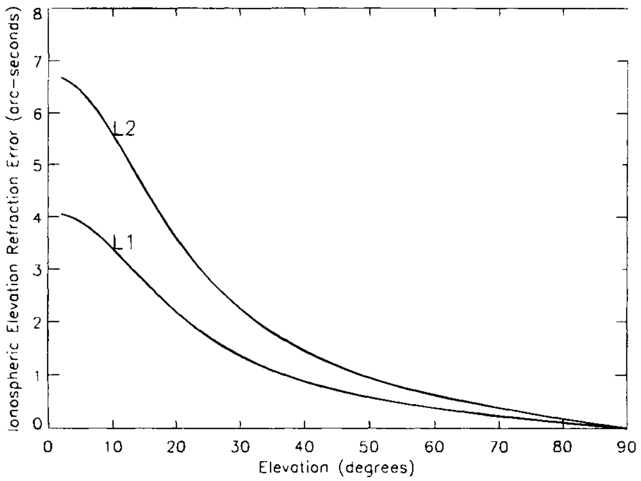


Fig. 3 Ionospheric refraction vs elevation for both L_1 and L_2.

Typical values of elevation refraction error for GPS at L_1 and at L_2, for a TEC of 10^{18} (el/m^2) column are given in Fig. 3.

The refraction, or radio wave bending illustrated in Fig. 3, is generally not a problem for GPS, because the user does not attempt to use GPS satellites at elevation angles lower than approximately 5 deg, due to other effects at low elevation angles, including antenna multipath, and tropospheric delay effects that increase greatly at low elevation angles. Errors in the azimuth of GPS radio signals transmitted through the ionosphere also can occur. They depend upon azimuthal gradients in TEC that generally are smaller than vertical gradients, and they can be neglected in most practical cases.

H. Distortion of Pulse Waveforms

Two characteristics of the ionosphere can produce distortion of pulses of rf energy propagated through it. The GPS signals consist of spread spectrum pseudorandom noise, having bandwidths of approximately 2 MHz and 20 MHz for the C/A and the P codes, respectively. The ionosphere can produce dispersion of the spread spectrum signals from GPS, but this effect is very small. The dispersion, or differential time delay caused by the normal ionosphere, as derived by Millman,[8] produces a difference in pulse arrival time across a bandwidth, Δf, of

$$\Delta t = [(80.6 \times \Delta f)/cf^3] \times \text{TEC}, \ldots, \text{seconds} \tag{19}$$

where c is the velocity of light in m/s, f and Δf are expressed in Hz, and TEC is in el/m^2 column. The dispersive term for pulse distortion is thus proportional to TEC. When the difference in group delay time across the bandwidth of the pulse is the same magnitude as the width of the pulse, it will be significantly disturbed by the ionosphere. The dispersion across the 20-MHz GPS bandwidth is normally small and can be ignored.

Chapter 13

Tropospheric Effects on GPS

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089

I. Tropospheric Effects

A. Introduction

THIS chapter discusses the effects of the troposphere on the GPS *L*-band signals and the resulting effect on GPS positioning. The specific effects discussed include tropospheric attenuation, scintillation, and delay. To be precise, the term tropospheric used in this chapter is somewhat of a misnomer because roughly 25% of the delay effect is caused by atmospheric gases above the troposphere, specifically gases in the tropopause and stratosphere as shown in Fig. 1. The troposphere produces attenuation effects that are generally below 0.5 dB and delay effects on the order of 2–25 m. These effects vary with elevation angle because lower elevation angles produce a longer path length through the troposphere and also vary with the detailed atmospheric gas density profile vs altitude.

1. *Atmospheric Constituents and Profile*

The atmosphere consists of dry gases and water vapor. The wet and dry constituents of the atmosphere affect the propagation delay of the radio frequency signals quite differently, and these constituents have different pressure profiles. Water vapor is confined to the troposphere and generally exists only below altitudes of 12 km above sea level, and most of the water vapor is below 4 km. Water vapor density varies widely with position and time and is much more difficult to predict than the dry atmosphere. For example, significant changes in water vapor can occur over tens of km and hours of time. Fortunately, however, water vapor effects represent only a relatively small fraction (roughly 1/10) of the total. Because the water vapor content is highly nonuniform, a total average fractional volume is not a very meaningful method of description of the water vapor content. Instead, a measure of water vapor often employed is the total amount of water in a vertical column of air one square cm in area:

Copyright © 1994 by the author. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Released to AIAA to publish in all forms.

*Ph.D., Chairman of the Board.

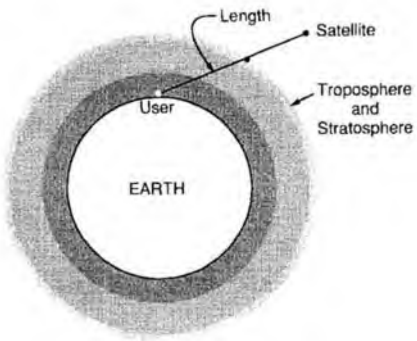


Fig. 1 Propagation of the GPS signal through the troposphere and stratosphere produces attenuation, delay and scintillation effects. The tropopause is the boundary between the troposphere and the higher altitude stratosphere.

$$W = \int_0^\infty \rho_w(h)\mathrm{d}h \text{ g}$$

where $\rho_w(h)$ is the water vapor density in g/cm^3. The total integrated water content W can vary enormously with position and ranges from 0.01 to 7.5 g, from the polar region to the tropics.

On the other hand, the dry atmosphere is relatively uniform in its constituents. Typical dry gas constituents have molar weights and fractional densities,* as shown in Table 1. At the GPS frequencies, oxygen is the dominant source of attenuation.

Table 1 Molar weights and approximate fractional volumes[a] of the major constituents of dry air[1]

Constituent	Molar weight, kg/kmol	Fractional volume unitless
N_2	28.0134	0.78084
O_2	31.9988	0.209476
Ar	39.948	0.00934
CO_2	44.00995	0.000314
Ne	20.183	0.00001818
He	4.0026	0.00000524
Kr	83.30	0.00000114
X_e	131.30	0.000000087

[a] For reference, at sea level and at 100% humidity, the water vapor occupies roughly 1.7% by volume, but note that percent humidity varies considerably with both time and position.[2]

* The fractional volume of CO_2 is the only major constituent of dry air that shows a significant variation, and it shows an annual variation of 6 ppm out of approximately 300 ppm and is increasing at a rate of 1.7 ppm/year.[3] However, for our purposes, these effects are negligible.

The temperature profile of a highly simplified model of the atmosphere in Fig. 2 shows that the temperature generally decreases with altitude at a constant lapse rate of −5 to −7°C per km of altitude increase from sea level up to the tropopause. At higher latitudes in winter and at nighttime, there is sometimes a temperature inversion layer in the 0.5–2 km region before the constant lapse rate of the troposphere begins. The tropopause is a region of approximately constant temperature and has an altitude of 8–12 km in the winter, or 10–12 km during the summer. The height of the tropopause has a downward slope from the equator toward the poles, and there is a small discontinuity in the tropopause height just above and below the equator. In the tropopause, the temperature rate of change decreases to zero. In the lower stratosphere (just above the tropopause), the rate of change of temperature gradually reverses to a slight +1°C to 2°C/km, and this relatively slow, nonuniform increase continues through the stratosphere up to an altitude of approximately 50 km. At this altitude, the stratopause, the temperature rate begins to reverse, and the temperature is roughly 0°C.[3,4] In the region directly above the stratopause; namely, the mesosphere, the temperature again decreases until it reaches approximately −90°C at the mesopause, which has a height of approximately 90 km. However, the atmospheric pressure in the mesosphere is so small (0.02–1 mb) as to be inconsequential for purposes here.

The atmospheric pressure at sea level is roughly 1013 mb and decreases with altitude to approximately 200–350 mb at the tropopause at the pole, and 70–150

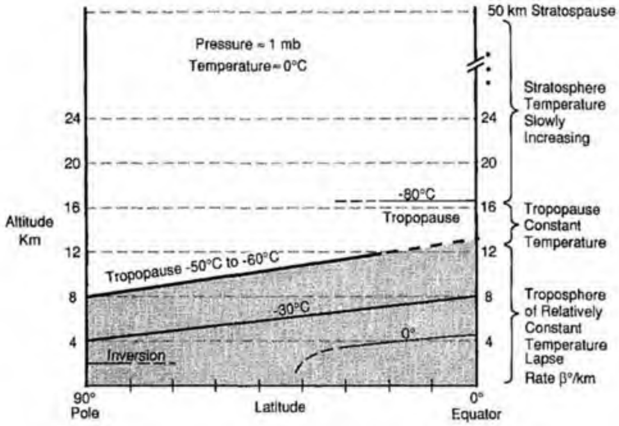

50 km Stratospause
Pressure ≈ 1 mb
Temperature ≈ 0°C
Stratosphere Temperature Slowly Increasing
24
20
16
12
8
4
24
20
16
12
8
4
Altitude Km
-80°C
Tropopause
Tropopause Constant Temperature
Tropopause -50°C to -60°C
-30°C
0°
Inversion
Troposphere of Relatively Constant Temperature Lapse Rate β°/km
90° Pole
Latitude
0° Equator

Fig. 2 Simplified model of the isothermal lines during the winter. Below the tropopause, the isotherms are approximately equally spaced at a temperature lapse rate of approximately 5–7°C/km decrease in temperature for every km of altitude increase. At the tropopause, the temperature lapse rate equals zero and then gradually reverses to a small rate of increase of approximately 1°–2°C/km in the stratosphere. Note that the height of the tropopause is roughly 4–8 km lower at the pole than near the equator. Thus, the tropopause has a downward slope toward the poles. During the summer, the tropopause is approximately 2 km higher in altitude at the poles but remains in the range of 17 km at the equator. At the polar regions, there is also an arctic temperature inversion at approximately 2 km, where the temperature lapse rate reverses.[3,4]

mb at the equator. Pressure decreases further to only 30 mb at approximately an altitude of 24 km at the pole and is only 1 mb at the stratopause. One atmosphere of pressure is defined as 1013.25 mb. The water vapor content lies primarily in the region below 4 km.

B. Atmospheric Attenuation

Atmospheric attenuation in the 1–2 GHz frequency band is dominated by oxygen attenuation, but even this effect normally is small. The attenuation is on the order of 0.035 dB for a satellite at zenith.[5,6] However, it can be ten times larger (in dB) at low elevation angles. The effects of water vapor, rain, and nitrogen attenuation at frequencies in the GPS frequency bands are negligible.[3]

Oxygen attenuation $A(E)$ in dB for the 1.5 GHz frequency range is approximately 0.035 dB at zenith ($E = 90$ deg) and varies with elevation angle E in proportion to the tropospheric path length L (obliquity factor or mapping function). If the troposphere is modeled by a simple uniform spherical shell of height h_m above the Earth, as shown in Fig. 3, then the length of the path L varies with elevation angle E, as shown in the fig. Thus, $A(E)$ has the following approximate value:

$$A(E) \cong \frac{2A(90\text{ deg})(1 + a/2)}{\sin E + \sqrt{\sin^2 E + 2a + a^2}}$$
$$\cong \begin{cases} \dfrac{2A(90\text{ deg})}{\sin E + 0.043}\text{ dB for small } E \text{ but } >3\text{ deg} \\[2ex] \dfrac{A(90\text{ deg})}{\sin E}\text{ dB for } E > 10\text{ deg} \end{cases} \qquad (1)$$

where $a = h_m/R_e << 1$ and h_o is the equivalent height for oxygen $h_m = 6$ km, and R_e is the Earth radius $R_e \cong 6378$ km.[7]

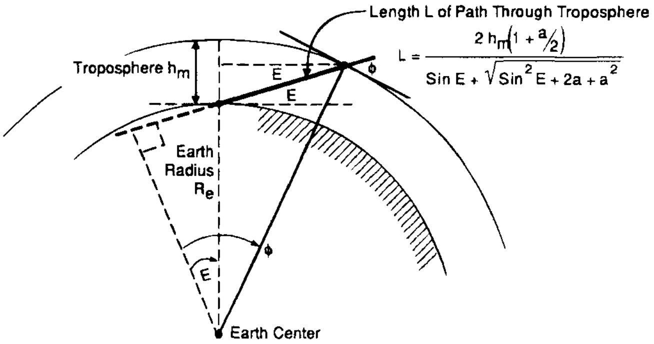

Length L of Path Through Troposphere
$L = \frac{2\,h_m(1 + a/2)}{\text{Sin } E + \sqrt{\text{Sin}^2 E + 2a + a^2}}$
Troposphere h_m
E
φ
E
Earth Radius R_e
φ
E
Earth Center

Fig. 3 **Path length L through a uniform shell troposphere at elevation angle E where $\cos E = (1 + a)\cos \phi$ where $a \triangleq h_m/R_e$.**

The attenuation of Eq. (1) is plotted in Fig. 4 for elevation angles $E > 3$ deg. This expression for $A(E)$ has assumed a spherical troposphere symmetrical in azimuth and uniform in density up to the equivalent height h_o. Note, if the troposphere were simply a planar layer with thickness h over a flat Earth, the length would simply be $h_0 csc\ E$. Near the horizon; e.g., below 3 deg, the uniform spherical model of Fig. 3 is no longer accurate, and neither Eq. (1) nor Fig. 4 should be used. The complexity of the atmosphere at low elevation angles is discussed later in this chapter. Nonetheless, note from Fig. 4 that the attenuation at 5 deg elevation angles is 0.38 dB, which is approximately ten times the zenith attenuation of 0.035 dB.

The mapping or obliquity function [i.e., the ratio $A(E)/A(90\text{ deg})$] at low elevation angles for the troposphere is significantly larger than that for the ionosphere (10 compared to 3) because the troposphere extends down to the surface of the Earth. Thus, at low elevation angles, the ray path to the satellite penetrates the lower troposphere in a more nearly horizontal direction than it does at the ionosphere, which exists only above approximately 60 km. In practice, we should avoid using GPS satellites below approximately 5-deg elevation not only because of the lower signal levels associated with tropospheric attenuation, but because of larger uncertainties in tropospheric and ionospheric delay and greater scintillation effects caused by both the troposphere and the ionosphere (discussed later in this chapter and discussed in the previous chapter for the ionosphere). Furthermore, potential signal multipath, refraction, and receiving antenna gain roll-off effects may be magnified at low elevation angles.

C. Rainfall Attenuation

For a frequency of 2 GHz, the attenuation even for dense, 100 mm/h rainfall, is less than 0.01 dB/km; thus, it has a very small effect. Rainfall attenuation below 2 GHz is even less; thus, rain attenuation is of little consequence in the frequency bands of interest for GPS, 1.57542 GHz and 1.2276 GHz.[5]

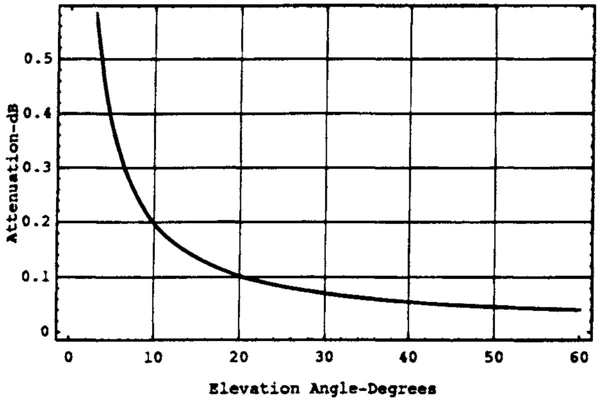


Fig. 4 Atmospheric attenuation vs elevation in degrees. Near the horizon at $E \leq 3$ deg, the model is no longer accurate.

D. Tropospheric Scintillation

Tropospheric scintillation is caused by irregularities and turbulence in the atmospheric refractive index primarily in the first few kilometers above the ground. The satellite–Earth propagation link through the troposphere is affected by a combination of random absorption and scattering from a continuum of signal paths that, in turn, cause random amplitude and phase scintillations in the received waveform. The scintillation effect varies with time and is dependent upon frequency, elevation angle, and weather conditions, especially dense clouds. At elevation angles above 10 deg, the predominant effect is forward scattering caused by atmospheric turbulence. At GPS frequencies, these effects are generally relatively small except for a small fraction of the time and at low elevation angles.

A received carrier from a satellite generally has the form $A(t)\sin(\omega t + \phi)$. Define the scintillation intensity $x(t)$ as the log of the amplitude ratio $A(t)/\overline{A(t)}$:

$$x(t) = 20 \log_{10}[A(t)/\overline{A(t)}] \text{ dB} \tag{2}$$

where $\overline{A(t)}$ is the mean (short-term) amplitude of the signal. The probability density of $x(t)$ in dB in the short term is Gaussian and has variance σ_x^2 (see Boithias,[5] Ippolito,[8] and Moulsey and Vilar[9]). Thus, $A(t)$ has lognormal statistics. Experimental measurements of received carrier amplitude[8,9] have shown that the statistics of amplitude scintillation $A(t)$, although well represented by a lognormal distribution (Gaussian in dB representation with rms value σ_x) in the short term are not truly stationary and, over a longer observation interval, have significant fluctuations in the σ_x parameter with time caused, for example, by changes in weather. The appearance of large irregularly shaped cumulus clouds in the satellite–user path can change the value of σ_x. In observations taken while a large cumulus cloud passed through the antenna beam, the value of σ_x increased by a factor of six (see Ippolito[8]), although when scaled to L-band these effects are still small.

Moulsey and Vilar[9] represent $A(t)$ with a conditional lognormal distribution. Thus the probability of $x(t)$ conditional on a given σ_x is then normal with a variance σ_x^2 and is then as follows:

$$p(x \mid \sigma_x) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp - \left(\frac{x^2}{2\sigma_x^2}\right) \tag{3}$$

The rms value σ_x in dB is itself a random variable with a mean σ_m, and its long-term fluctuations have a probability density that is also lognormal:

$$p(\sigma_x) = \frac{1}{\sigma_\sigma \sigma_x \sqrt{2\pi}} \exp - \left[\frac{(\log \sigma_x - \log \sigma_m)^2}{2\sigma_\sigma^2}\right] \tag{4}$$

where σ_σ^2 is the variance of this distribution. Thus $p(x|\sigma_x)$ can be viewed as a short-term statistic wherein σ_x itself has a distribution for the long term. The rms value of x in the long term is defined as σ_{xm}, which can be evaluated using $p(x) = \int p(x|\sigma_x)p(\sigma_x)\mathrm{d}\sigma_x$ and found to be $\sigma_{xm} = \sigma_m \exp(\sigma_\sigma^2/4)$ dB.

For a value $\sigma_\sigma = 1$, which seems to match measurements by Moulsey and Vilar,[9] the value of $\sigma_{xm} = 1.28\, \sigma_m$. The interpretation of this result is that although the long-term mean value of σ_x is σ_m, and σ_m may be small; e.g., on the order of 0.259 dB, the variance $\sigma_\sigma^2 = 1$ can lead to momentary periods of much larger scintillation.

The CCIR (1982) has given an expression for the long-term mean value of σ_x, namely, σ_m. For small antennas such as omnidirectional GPS antennas, the CCIR expression for the long-term rms amplitude scintillation varies with frequency and elevation angle as follows:

$$\sigma_m = 0.025 f^{7/12} (\csc E)^{-0.85} \text{ dB} \tag{5}$$

where f is in GHz, and the elevation angle is E. (Larger antennas, not of importance for most GPS applications, produce an antenna averaging effect that decreases the value of σ_m). For $L_1 = 1.57542$ GHz we have the following:

$$\sigma_m = 0.0326 (\csc E)^{-0.85} \text{ dB} \tag{5a}$$

and $\sigma_m = 0.259$ dB at $E = 5$ deg. Over the long term, the value of σ_x can vary substantially. If the mean σ_m of σ_x is taken as $\sigma_m = 0.259$ corresponding to the (worst case) 5 deg elevation angle for the CCIR model, and $\sigma_m = 0.081$ dB for 20-deg elevation angle, the cumulative distribution of σ_x is then as shown in Fig. 5. Thus, roughly 10% of the time the rms scintillation σ_x has a value of 0.9 dB and 0.3 dB for these two elevation angles, respectively, for this model. Thus, for low elevation angles and small fractions of time, tropospheric scintillation can be significant, but otherwise it is quite small.

II. Tropospheric Delay

The signal received from a GPS satellite is refracted by the atmosphere as it travels to the user on or near the Earth's surface. The atmospheric refraction causes a delay that depends upon the actual path (slightly curved) of the ray and

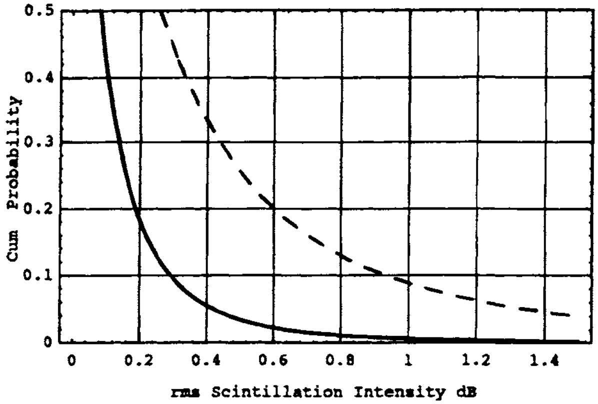


Fig. 5 Cumulative probability for the rms tropospheric scintillation σ_x dB. For this example, the mean value of $\sigma_m = 0.259$ dB for small 5 deg elevation angles (dashed curve) and $\sigma_m = 0.081$ dB for 20 deg elevation angle solid curve. For this lognormal model and a cumulative probability of 10%, the rms scintillation intensity σ_x is equal to 0.9 dB and 0.3 dB, respectively, for 5-deg and 20-deg elevation angles.

the refractive index of the gases along that path. For an atmosphere symmetric in azimuth about the user antenna, the delay depends only upon the vertical profile of the atmosphere and the elevation angle to the satellite.

A. Path Length and Delay

There are two major delay effects of the troposphere. The first and larger effect is a dry atmosphere excess delay caused primarily by N_2 and O_2. The dry atmosphere zenith excess delay corresponds to approximately 2.3 m and varies with local temperature and atmospheric pressure in a reasonably predictable manner.* The dry atmosphere effect varies by less than 1% in a few hours. The second effect—the wet atmosphere or water vapor effect—is generally smaller, 1–80 cm at zenith; i.e., perhaps $^1/_{10}$ the size of the dry atmosphere delay effect, but it varies markedly, 10–20%, in a few hours and is less predictable even with surface humidity measurements. Dual frequency 22, 31 GHz radiometer measurements can be made to make more precise predictions, but these measurements are fairly complicated and not feasible for most navigation applications.

The tropospheric delay is caused by the larger refractive index n ($n > 1$) of atmospheric gases than that of free space ($n = 1$), which causes the speed of light (group velocity) in the medium to decrease below its free space value c. Thus, a ray with an infinitesimal path length $\mathrm{d}s$ that travels through a medium with a refractive index n has a time delay $\mathrm{d}\tau = n\,\mathrm{d}s/c$ and has an equivalent distance $c\,\mathrm{d}\tau = n\,\mathrm{d}s$. In this section, delays are measured in meters rather than time, and the c is omitted after this paragraph. The difference between the actual total path delay $S = c\tau = \int n(s) \times 10^{-6}\,\mathrm{d}s$ and the geometrical path distance S_g is the excess tropospheric delay Δ. The difference between the actual refractive index and unity is of the order of magnitude of $n-1 \cong 2.7 \times 10^{-4}$ at sea level and varies with altitude, latitude, and various meteorological conditions.

At the L-band frequencies of GPS signals, the refractive index is essentially constant with frequency, and hence, is nondispersive; i.e., $\mathrm{d}n/\mathrm{d}f = 0$, and thus, the group velocity and phase velocity are the same. As shown in the previous chapter, this equivalence is not true in the ionosphere.

The spatially varying refractive index causes the signal path to have a slight curvature with respect to the geometric straight line path. Thus, the actual satellite-to-user path is a slightly curved path, as shown in Fig. 6 where the path curvature has been greatly exaggerated for clarity. The total length of the actual path from P_1 to P_2 in m is as follows:

$$S = \int_{\text{Actual}} \mathrm{d}s \tag{6}$$

and is clearly longer than the straight line geometric path S_g (at elevation angles less than 90 deg) that the ray would take in a vacuum. However, it is well known from Fermat's principle[5,10] that the actual ray path is that path which minimizes the total delay from P_1 to P_2 (omitting c from $c\tau$ from now on):

* In the frequency range below 15 GHz, both the wet and dry components have an effect. At optical frequencies, the whole effect is attributable to the dry component.

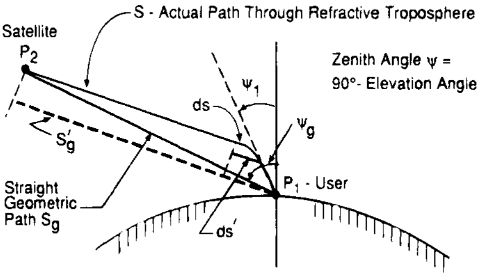


Fig. 6 Actual slightly curved path S and straight geometric paths through the refractive atmosphere. The difference in the actual path distance vs the geometric path distance is $\Delta_g \triangleq S - S_g \cong S - S_g{}'$ where $\Delta_g \triangleq \int \mathrm{d}s - \int \mathrm{d}s'$.

$$\tau = \int_{\text{Actual}} n(s)\,\mathrm{d}s \tag{7}$$

The actual (curved) path from P_1 proceeds along a path which is initially closer to the zenith so that it passes through the region of high dielectric constant more rapidly. Thus, the actual curved path length S is longer than the straight line geometric path length (a rectilinear chord) S_g in m:

$$S_g \triangleq \int_{\text{geo}} \mathrm{d}s \quad \text{and} \quad \tau_{\text{geo}} = \int_{\text{geo}} n(s)\,\mathrm{d}s \tag{8}$$

although it is shorter in time delay. The notation τ_{geo} represents the path delay which would be accumulated along the geometric rectilinear chord. The integral path notation geo defines the geometric straight line path. That is, delay τ is shorter than the delay τ_{geo} of the straight line path. The quantity of primary interest is the excess delay caused by the atmosphere $\Delta = \tau - S_g$. The difference between the length of the actual curved path and the straight line path is typically quite small $S - S_g < 0.1$ m[11] except at low elevation angles.

If point s along the ray path is expressed in polar coordinates, then $s = re^{j\phi}$ (two dimensions), where the center of the coordinate system is at the Earth center, then the total delay [Eq. (7)], can be written as follows:

$$\tau = \int_{\text{Actual}} n(s)\,\mathrm{d}s = \int_{\text{Actual}} n(r,\phi)\sqrt{1 + (r\phi')^2}\;\mathrm{d}r \tag{9}$$

where $\mathrm{d}s = \mathrm{d}r\sqrt{1 + (r\phi')^2}$ and $\phi' \triangleq \mathrm{d}\phi/\mathrm{d}r$ for the path. Assume that $n(r,\phi) = n(r)$; i.e., the troposphere is spherically symmetric about the Earth's center. Fermat's principle and the calculus of variations then gives a minimum τ for the path and results in the following relationship:

$$\frac{n(r)r^2\phi'(r)}{\sqrt{1 + (r\phi')^2}} = \text{constant, or } n(r)r\sin\psi(r) = n(0)R_e\sin\psi(0) = \text{constant} \tag{10}$$

which is Snell's Law, and where $\psi(r)$ is the angle between the Earth's center radial vector at (r,ϕ), and the path tangent at height h above the Earth (see Fig.

7). At the user point P_1 on the Earth's surface, the radial distance $r = R_e$, the Earth radius, and the angle $\psi(0)$ is the zenith angle at the user $\psi(0) = 90$ deg $- E$, where E is the elevation angle. Another method of arriving at the same result is to assume a concentric set of spherical shells of height Δr and constant refractive index n_i, as shown in Fig. 7. Thus, we have Snell's law for spherical shells as follows:

$$r_i n_i \sin \psi_i = r_1 n_1 \sin \psi_1 = n(0) R_e \sin \psi(0) = \text{constant} \tag{10a}$$

In summary, the curved path, although physically slightly longer, has a slightly shorter total delay than would a path traveling in a straight line through the troposphere. Thus, the difference between the actual curved delay and straight line geometric distance (rectilinear chord) is as follows:

$$\Delta = \tau - S_g = \int_{\text{Actual}} n(s)\, ds - \int_{\text{geo}} ds \tag{11}$$

The second integral is the delay for the straight line (neglecting relativity) path that the ray would take in a vacuum. This difference can be rewritten as follows:

$$\Delta = \int_{\text{Actual}} (n - 1)\, ds + \left[\int_{\text{Actual}} ds - \int_{\text{geo}} ds \right] = \int_{\text{Actual}} (n - 1)\, ds + \Delta_g \tag{12}$$

where the second term in brackets Δ_g is the difference between the curved and

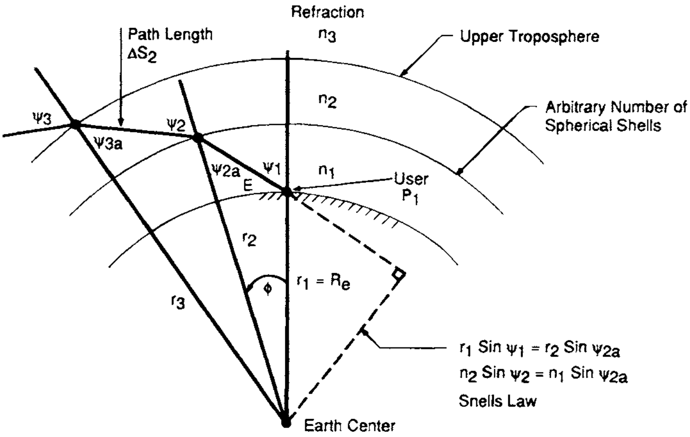


Fig. 7 Diagram of a ray path through sequential spherical shells of uniform refractive index n_i with base shell radi r_i and heights $h_i = r_i - R_e$, respectively. Snells law for spherical shells is $r_1 n_1 \sin \psi_1 = r_2 n_2 \sin \psi_2$. For planar boundaries Snells law has the familar form $n_2 \sin \psi_2 = n_1 \sin \psi_{2a}$. The same relationships hold for an arbitrary number of uniform shells of width Δr. The length of each path within a shell of infinitesimal width dh is related to the shell width by the ratio

$$ds/da = R_e(1 + a\, n(a))/(2a\, n(a) + a^2 n^2(a) + \sin^2 E)^{1/2}$$

where $a \triangleq h/R_e$, $n(a)$ is the refractive index; and a is the normalized height.

free-space paths $S - S_g \overset{\Delta}{=} \Delta_g$. Ray-tracing calculations comparing excess delay Δ for a straight line path vs. a curved path from Snell's law have been calculated by Janes et al.[7] for the U.S. Standard atmosphere for various latitudes and elevation angles. The error caused by the neglect of path curvature is less than 3 mm for $E \geq 20$ deg; 2 cm for $E = 10$ deg; and increases to 17 cm at $E = 5$ deg (primarily from dry gases). The initial elevation angle difference between the straight line path and the curved path is on the order of seconds of arc, except for low elevation angles. The actual difference between the actual zenith angle $\psi(0)$ and the geometric straight line zenith angle ψ_g is difficult to compute (we must integrate $\int \tan \psi \, dn/n = \psi_g - \psi(0)$). However, a loose upper bound is easily obtained as follows:

$$\Delta\psi = \psi_g - \psi(0) < N(0) \times 10^{-6} \tan \psi(0) \tag{13}$$

by integrating over a single shell of constant refractive index $N(0)$ (see also Ref. 9). A more precise formula (radio ranging standard formula) from Saastamoinan[11] *is as follows:*

$$\Delta\psi \text{ in seconds of arc} \cong \frac{16.0 \tan \psi_g}{T}\left(P + \frac{4800e}{T}\right) - 0.07(\tan^3 \psi_g + \tan \psi_g)\left(\frac{P}{1000}\right) \tag{14}$$

where P and e are the atmosphere pressure and the partial pressure of water vapor in millibars, respectively, and T is the absolute temperature in °Kelvin. The quantity $\Delta\psi$ is the angle of refraction. The angle $\Delta\psi$ is generally quite small, generally $\Delta\psi < 0.1$ deg.

In general, the estimation models for the actual troposphere excess delay from point P_1 to point P_2 is as follows:

$$\Delta = (10^{-6}) \int_{\text{Actual}} N ds + \Delta_g \tag{15}$$

where the refractivity $N \overset{\Delta}{=} (n - 1) \times 10^6$ has both dry and wet terms. The excess delay Δ can be approximated as the sum of wet and dry delay effects in a simplified form similar to the following:

$$\Delta = [\Delta_{zd} m_d(E) + \Delta_{zw} m_w(E)] \tag{16}$$

where Δ represents the excess tropospheric delay; Δ_{zd} represents the dry zenith delay; Δ_{zw} represents the wet zenith delay; and $m_d(E)$ and $m_w(E)$ are the dry and wet mapping functions (obliquity factor) that magnify the tropospheric delay as the elevation angle E decreases. In general, $m_d(E)$ and $m_w(E)$ are really functions of the atmospheric profile as well as E. However, expressing them as a function of E alone is a useful approximation.

To analyze Δ further, it is first necessary to investigate the refractivity profiles of the troposphere with altitude. These results can then be employed to compute the zenith delay terms Δ_{zd} and Δ_{zw}. Finally, models can be developed for the mapping functions $m_d(E)$ and $m_w(E)$. We must be careful to understand exactly what is meant by the wet and dry terms because the actual path of the ray is dependent on all contributions to the index of refraction. As shown below, it is often convenient to break up the excess delay into a hydrostatic component Δ_h, which is dependent only upon the total pressure (and not the mix ratio of dry and wet air) and a wet component Δ_w.

B. Tropospheric Refraction Versus Pressure and Temperature

The general empirical expression for refractivity of nonideal gases including water vapor to be used here is as follows (see Davis et al.[1] and Thayer[12]).

$$\begin{aligned} N = (n-1)10^6 &= k_1(P_d/T)Z_d^{-1} + k_2(e/T)Z_w^{-1} + k_3(e/T^2)Z_w^{-1} \\ &= 77.604(P_d/T)Z_d^{-1} + 64.79(e/T)Z_w^{-1} + 377600(e/T^2)Z_w^{-1} \\ &= 77.604(P_d/T)Z_d^{-1} + (e/TZ_w)(64.79 + 377600/T) \qquad (17) \end{aligned}$$

where P_d and e are the partial pressure of dry air and partial pressure of water vapor both in millibars; T is the temperature in °Kelvin; and the compressibility Z factors correct for the small departures of the moist atmosphere from an ideal gas. Ideal gases obey the relationship $PV = RT$; nonideal gases operate as $PV = ZRT$ where R is the universal gas constant. The experimentally derived values used here are: $k_1 = 77.604 \pm 0.0124$; $k_2 = 64.79 \pm 10$; $k_3 = 377600 \pm 3000$. See Table 2 for various values of k_i.

The inverse compressibility expressions are as follows[1]:

$$Z_d^{-1} = 1 + P_d[57.97 \times 10^{-8}(1 + 0.52/T) - 9.4611 \times 10^{-4}\, T_c/T^2]$$

$$Z_w^{-1} = 1 + 1650\,(e/T^3)[1 - 0.01317T_c + 1.75 \times 10^{-4}T_c^2 + 1.44 \times 10^{-6}T_c^3] \qquad (18)$$

where T_c is temperature in °Celsius; T is in °Kelvin; and P_d is the dry pressure in millibars. The molar weight of dry air is $M_d = 28.9644$ kg/kmol, and for water $M_w = 18.0152$ kg/kmol. The ratio $M_w/M_d = 0.621977$.

The task at hand is to define models for the pressure and temperature vs. altitude that will yield models for N vs altitude h using Eqs. (17) and (18). The first term in Eq. (17) is referred to as the dry term, and the second two terms as the wet terms. Typically, the dry term represents perhaps 90% of the total effect and is relatively well predicted by surface pressure measurements. The wet terms, although small, are not easily predicted, and the water vapor effects vary considerably with position, elevation angle, and time. The pressure P and density ρ for the dry and wet air are related by the gas law as follows:

$$P_d = \rho_d \frac{R}{M_d} TZ_d, \qquad P_w = e = \rho_w \frac{R}{M_w} TZ_w \qquad (19)$$

where ρ_d and ρ_w are the dry and wet densities, respectively.

Thus, the refractivity N can be expressed in terms of the densities by substituting Eq. (19) into the first two terms of Eq. (17) as follows:

$$\begin{aligned} N &= k_1 \frac{R\rho_d}{M_d} + k_2 \frac{R\rho_w}{M_w} + k_3 \frac{e}{T^2} Z_w^{-1} \\ &= k_1 \frac{R\rho}{M_d} + \left(k_2 - k_1 \frac{M_w}{M_d}\right) \frac{e}{T} Z_w^{-1} + k_3(e/T^2)Z_w^{-1} \end{aligned}$$

Table 2 Constants used in the refraction Eq. (17) by various sources for microwave frequencies. In this chapter, the Thayer values are employed.

Source	k_1(°K/mb)	k_2(°K/mb)	k_3(°K^2/mb)
Thayer[12] (with Davis et al.,[1] uncertainty)	77.604 ± .014	64.79 ± 10	377600 ± 3000
Hill,[13] theoretical	——	98 ± 1	358300 ± 300
Birmbaum and Chatterjee,[14] measurements	——	71.4 ± 5.8	374700 ± 2900
Boudouris,[15] measurements	——	72 ± 11	375000 ± 3000

$$= (77.604 \pm 0.014)\,\rho\,\frac{R}{M_d} + k_2'\,\frac{e}{T}\,Z_w^{-1} + k_3\,\frac{e}{T^2}\,Z_w^{-1}$$

$$= 22.276\,\rho + (16.5 \pm 10)\,\frac{e}{T}\,Z_w^{-1} + 377600\,\frac{e}{T^2}\,Z_w^{-1} \qquad (20a)$$

where $\rho = \rho_d + \rho_w$ is the total mass density, $k_2' = k_2 - k_1\,M_w/M_d$; and, of course, both the ρ_d, ρ_w densities and T vary with height h. Note that there is considerable uncertainty in the coefficient of the $(e/T)Z_w^{-1}$ term.[1] The uncertainty in the k_1 term is ±0.018%. Note that the first term in Eq. (20a) is dependent only on the total density ρ and is independent of the fractional water vapor content. The first term in the last version of Eq. (20) is termed the hydrostatic component. Total pressure is related to total density through the hydrostatic equation $\mathrm{d}P/\mathrm{d}h = -\rho(h)\,g(h)$, where $g(h)$ is the acceleration of gravity at the receiver antenna that also has a variation with height (only small variation for the relatively short range of heights for the troposphere). Generally $g(h)$ has very little variation with h, and a good approximation is obtained by replacing $g(h)$ by its weighted mean value over the range of integration $g_m = [\int\rho(h)g(h)\mathrm{d}h]/[\int\rho(h)\mathrm{d}h]$.

For this reason, we rewrite Eq. (20a) as the sum of hydrostatic and wet refractivity:

$$N = N_h + N_w \qquad (20b)$$

where the hydrostatic refractivity is $N_h = (k_1R/M_d)\,\rho = 22.276\,\rho$, and the wet refractivity is as follows:

$$N_w = \left[k_2 - k_1\,\frac{M_w}{M_d}\right]\frac{e}{T}\,Z_w^{-1} + k_3\,\frac{e}{T^2}\,Z_w^{-1}$$

The hydrostatic refractivity is sometimes referred to as the dry refractivity and differs only by the Z_d^{-1} coefficient.

1. Temperature, Pressure, Refraction Index Models

The variation of the refractivity of the atmosphere vs altitude is exceedingly complex, and the models for refractive index are in the end based on fitting coefficients to match extensive sets of measured data. Nevertheless, it is instructive to review fundamental thermodynamics and ideal gas laws because some of the more complex models bear a close resemblance to these idealized models.

a. Troposphere–ideal gas models—adiabatic process. First consider an atmosphere of a single ideal gas with a pressure vs. altitude $P(h)$. At the lower altitudes within the troposphere, there is continual upward and downward motion of the air masses. Assume as an approximation that the process is assumed to be reversible and adiabatic (no heat entering or leaving the process). Also, assume that the gases are ideal* $PV = RT$. For an ideal gas with constant pressure heat capacity C_p and constant volume heat capacity C_y where $C_p = C_v + R$, a reversible adiabatic process with no heat transfer yields the following:

$$C_v \, \mathrm{d}T + P \, \mathrm{d}V = 0 \tag{21a}$$

Because $P = RT/V$ and $R/C_v = \gamma - 1$ where $C_p/C_v = \gamma$ is defined as the ratio of heat capacities. By using $PV = RT$ and $R/C_v = \gamma - 1$, Eq. (21a) becomes the differential equations that follow:

$$\mathrm{d}T/T = -R/V \, \mathrm{d}V = -(\gamma - 1) \, \mathrm{d}V/V, \quad \text{or} \quad \mathrm{d}T/T = [(\gamma - 1)/\gamma] \, \mathrm{d}P/P \tag{21b}$$

Equation (21b) has the following solutions:

$$\left(\frac{T_2}{T_1}\right) = \left(\frac{V_1}{V_2}\right)^{\gamma - 1}, \quad \text{and} \quad \left(\frac{T_2}{T_1}\right) = \left(\frac{P_2}{P_1}\right)^{(\gamma - 1)/\gamma} \tag{21c}$$

For a diatomic gas such as O_2, the heat capacity ratio changes slowly with temperature and has an approximate value of $\gamma = 1.4$. For polyatomic gases such as H_2O and CO_2, the value of γ is usually less than 1.3. Define the gas density as $\rho = M/V$, where M is the molar mass, and V is the volume. The hydrostatic relationship for the pressure P is dependent on the height h of the air column. For a differential air column of height $\mathrm{d}h$ the hydrostatic pressure increases by $(\rho \mathrm{d}h)g \, \mathrm{d}P$ at the bottom of the differential element compared to the top by the differential weight of the air per unit area:

$$\mathrm{d}P = -\rho(h)g(h) \, \mathrm{d}h = -\frac{Mg}{V} \, \mathrm{d}h, \quad \text{and} \quad \frac{\mathrm{d}P}{P} = \frac{-Mg}{RT} \, \mathrm{d}h \tag{22}$$

Assume now that the acceleration of gravity g is constant. Thus, using Eqs. (21) and (22), we have the following:

$$\begin{aligned} \frac{\mathrm{d}T}{T} &= \left(\frac{\gamma - 1}{\gamma}\right)\frac{\mathrm{d}P}{P} = -\left(\frac{\gamma - 1}{\gamma}\right)\left(\frac{Mg}{RT}\right) \mathrm{d}h, \quad \text{or} \\ \frac{\mathrm{d}T}{\mathrm{d}h} &= -\left(\frac{\gamma - 1}{\gamma}\right)\frac{Mg}{R} \cong \text{constant} = \beta \end{aligned} \tag{23}$$

if γ is also constant with h over the region of interest. For $\gamma = 1.4$, the rate of change of temperature with altitude is then as follows:

* Note that real gases do not strictly obey the ideal gas law, and the van der Waals equation $[P + (a/v^2)]\,(v - b) = RT$ is a more accurate representation.[16] In this chapter, the effects of nonideal gases are taken into account by the compressibility factors Z_d and Z_w for the dry gases and water vapor.

$$\beta = -34.0866(\gamma - 1)/\gamma = -9.74 \text{ deg/km} \quad \text{for} \quad \gamma = 1.4$$

$$M = M_d = 28.9644$$

$$\beta = -7.87 \text{ deg/km} \quad \text{for} \quad \gamma = 1.3$$

Thus, the adiabatic assumption with constant C_v, C_p leads to a constant lapse rate $\mathrm{d}T/\mathrm{d}h = \beta = -[(\gamma - 1)/\gamma]$ Mg/R. See Table 3 for example measured values of β, which are on the order of $(-6 \pm 0.5$ deg)/km, and other useful physical constants.

For nitrogen/oxygen atmosphere with $\gamma = 1.4$, the pressure then varies with altitude h as follows:

$$P = P_0\left(1 + \frac{\beta h}{T_0}\right)^{\gamma/(\gamma-1)} = P_0\left(1 + \frac{\beta h}{T_0}\right)^{1.4/(1.4-1)} = P_0\left(1 + \frac{\beta h}{T_0}\right)^{3.5} \qquad (24)$$

where the ratio of heat capacities for diatomic gases is $\gamma = 1.4$, and T_0 is the temperature at sea level. The typical observed values of β are negative at approximately -5 to -7°C/km in the troposphere. For $\gamma = 1.3$, the exponent would be 4.33. Measurements give a ratio γ of heat capacitances, which corresponds to a somewhat higher exponent of $\gamma/(\gamma - 1) = 5.17$.[17]

We can also simply assume at the outset that the temperature has a constant lapse rate without the adiabatic assumption; i.e., $T = T_0 + \beta h$, where β is expressed in units of deg/km. Then, from the hydrostatic equation, we can integrate $\mathrm{d}P/P = -(Mg/RT)\,\mathrm{d}h$ for constant g over the range of heights $h = (0,h)$ to obtain the following:

$$\int_{P_0}^{P} \frac{\mathrm{d}P}{P} = \frac{-Mg}{R}\int_0^h \left(\frac{\mathrm{d}h}{T_0 + \beta h}\right) \quad \text{yields} \quad P = P_0\left(\frac{T_0 + \beta h}{T_0}\right)^{-Mg/R\beta} \qquad (25)$$

where the surface pressure $P = P_0$ at $h = 0$. Thus, because $N_h \sim P_d/T$, the hydrostatic refractivity N_h can be written in terms of the surface value of hydrostatic refractivity N_{do} with a power law dependence on h as follows:

$$N_h = N_{h0}\left(\frac{T_0 + \beta h}{T_0}\right)^{\mu} \qquad (26)$$

where $\mu \triangleq (-Mg/R\beta) - 1$. Thus, the exponent depends on the lapse rate β. If $\beta = -6.81$°C/km, then the exponent $-Mg/R\beta - 1 = 4.0$, a value shown later to be used by Hopfield[18,19] in a two quartic model of the troposphere pressure variation vs. altitude. Other more complex models[11,20] also use a power law model but only for the troposphere, not the stratosphere. (Note that at an altitude $h = -T_0/\beta$, the temperature would have fallen to absolute zero in this model if β remained constant for all h.) The rate of change of N_h at $h = 0$ for this model is then obtained by differentiating Eq. (26) to obtain the following[21]:

$$\mathrm{d}N_n/\mathrm{d}h = \beta\mu N_{h0}/T_0 = \frac{-N_{h0}Mg}{RT_0}\left(1 + \frac{R\beta}{Mg}\right) \quad \text{at} \quad h = 0$$

See Table 3 for typical values of $\mathrm{d}N_h/\mathrm{d}h$ (shown as $\mathrm{d}N/\mathrm{d}h$).

Table 3 Thermodynamic constants and typical values of refraction variables and height of tropopause[1,11,17]

Typical measured values of tropopause height and lapse rates

Location	Location, north latitude	Tropopause, h_T km	Lapse rate, β °K/km
Onsala, South Sweden	57°	10.5	−5.7
Ettelsburg, Germany	51°	9.6	−5.7
Haystack, East Massachusetts	43°	13.6	−5.6
Owens Valley, Southern California	37°	12.8	−5.6
Fort Davis, Southwest Texas	31°	13.4	−6.3

Typical values of key parameters at surface and tropopause (models 1,2,3)[20]

Zone	Radius, km	P_0 *mb*	*T*, °K	$(n-1)$ 10^6	$-(dN/dh)$ 10^6/km	β, °K/km
Tropical	6360 + 0	1010.60	299.85	265.72	24.82	−6.06
Tropical	6360 + 16.8	98.03	198.00	39.06	5.525	
Temperate	6380 + 0	1015.00	285.08	280.87	27.28	−6.45
Temperate	6380 + 10.4	245.33	218.00	88.78	11.28	
Arctic	6400 + 0	1020.00	252.50	318.67	56.96	−6.525
Arctic	6400 + 8.8	303.56	223.60	107.39	13.33	

g_m = weighted mean gravitational constant = $[\int \rho(h)g(h)dh]/[\int \rho(h)dh]$
= $9.8062(1 - 0.0026\cos 2\lambda - 0.00031\,\bar{h})ms^{-2}$ where $\bar{h}$ is the effective height above sea level of the center of gravity of the atmospheric column. Saastamoinen[20] has shown that $\bar{h} \cong 7.3 + 0.9\,h_0$ km where h_0 is the user height and λ is the user latitude.

g_{eff} = effective gravitational constant = $9.784\ (1 - 0.0026\cos 2\lambda - 0.00028h_0)ms^{-2}$ is the effective value of g for this $\bar{h}$. λ is the user latitude and h_0 is the user height.

1 atmosphere = 1013.25 millibars

M_d = mean molar mass of dry air 28.9644 kg/kmol

M_W = molar mass of water = 18.0152 kg/mol

1 kJ = 10^3 kg m^2/s^2 = 10^{10} erg

R = universal gas constant 8.31434 kJ/(kmol · °K)

$$\frac{M_d g}{R} = \frac{(28.9644)(9.784)}{8.314}\,\frac{^\circ\mathrm{K}}{\mathrm{km}} = 34.0866\,\frac{^\circ\mathrm{K}}{\mathrm{km}}.$$ For dry air at sea level $h = 0$ and $\lambda = 45$ deg.

0°C = 273.16°K

Key:

C_v, diatomic gas Heat capacity, constant volume ≈ $(5/2)R$ at room temperature.

C_p, diatomic gas Heat capacity, constant pressure ≈ $(7/2)R$; $C_p = C_v + 1$.

b. Stratosphere—isothermal model of pressure vs altitude. At higher altitudes from the top of the troposphere (termed the tropopause), at altitude h_T ($h_T \cong$ 12 km), and throughout the stratosphere to the stratopause, a constant temperature (isothermal) condition can be used as an approximation. Except in the tropospause itself, the temperature is not truly constant but varies more slowly than in the troposphere. For a region of ideal gases of constant temperature, the eq. $\mathrm{d}P/P = (Mg/RT)\,\mathrm{d}h$ can be integrated easily because T is a constant. This isothermal (constant temperature) condition leads to the exponential barometric height equation for pressure at altitudes above h_T:

$$P = P_0 \exp\left[\frac{-Mg(h - h_T)}{RT_T}\right] \tag{27}$$

where R is the universal gas constant, 8.31434 kJ/(kmol · °K); g is the effective Earth's free fall acceleration in the stratosphere; P_T is the pressure at the tropopause; T_T is the temperature at the tropopause; and M is the molar mass of the air ($M \approx M_d$ in the stratosphere) (see Table 3). Thus, for the isothermal model, pressure decreases exponentially with height.

2. *Zenith Delay for Dry Gases—Hydrostatic Component*

The exact variation of the refractive index vs altitude is critical if we need to determine the excess tropospheric delay vs elevation angle, especially at the lower elevation angles (large zenith angles). However, for a satellite at the user's zenith, the hydrostatic gas effect of the atmosphere is independent of the shape of the profile with height and the dry air/water vapor mix ratio. The differential dry delay error at zenith Δ_{dz} (for no water vapor) from Eq. (17) is as follows:

$$\mathrm{d}\Delta_{dz} = N_d(h) \times 10^{-6}\,\mathrm{d}h \quad \text{where} \quad N_d(h) = 77.624\, P_d/T \tag{28}$$

where $N_d(h)$ is the dry refractivity as a function of altitude h. In this analysis, it is more useful to work with the total hydrostatic delay (which is primarily the dry delay component) where the total hydrostatic refractivity from Eq. (20) is $N_h(h) = 22.276\,\rho(h)$, and the zenith hydrostatic delay is $\Delta_{hz} = 22.276 \int \rho(h)\mathrm{d}h$. Note that the surface pressure P_0 is related to the density ρ by the hydrostatic equation $P_0 = \int \rho(h)g(h)\mathrm{d}h \cong g_{\mathrm{eff}} \int \rho(h)\mathrm{d}h$ where $g_{\mathrm{eff}} = \int \rho(h)g(h)\mathrm{d}h/\int \rho(h)\mathrm{d}h$ is the effective gravitational constant and is approximated in Table 3. Thus, the zenith excess hydrostatic delay [see Eq. (20)] is as follows:

$$\Delta_{hz} = 22.276 \times 10^{-3}\, P_0/g_{\mathrm{eff}} = 2.2768 \times 10^{-3} P_0/G \text{ m} \tag{29}$$

where we define $G \triangleq (1 - 0.002626 \cos 2\lambda - 0.00028 h_0) = g_{\mathrm{eff}}/9.784$, and P_0 is in millibars.

Water vapor does not conform to the hydrostatic equation because water vapor is subject to myriad effects including condensation and has wide variations in concentration unlike the dry gases that are uniformly mixed. Thus, there is no simple, highly accurate estimate of the wet delay related to the surface vapor pressure. The use of water vapor radiometers is generally impractical for most GPS applications; however, where they are available, Elgered[3] has developed expressions to estimate delay from these measurements to within a few mm for zenith delay.

III. Empirical Models of the Troposphere

A. Saastamoinen Total Delay Model

In a series of papers in 1972 and 1973, Saastamoinen[11,20] presented one of the first models of the refraction of the troposphere that estimates delay vs elevation angle E. In this model, the dry pressure is modeled using the constant lapse rate model for the troposphere and an isothermal model above the tropopause. The vertical gradient of temperature is $T = T_0 + \beta(r - r_0)$, and the resulting pressure profile is $P = P_0(T/T_0)^{-Mg/R\beta}$ where r is the radius from the Earth center ($r = R_e + h$) and r_0 is the user radius (usually $r_0 = R_e$, the Earth radius), and T_0 is the user temperature. The radius r ranges in value from r_0 to r_T which represents the radius to the tropopause. The corresponding dry refractivity, as discussed previously, is then $n - 1 = (n_0 - 1)(T/T_0)^{\mu}$ where $\mu = -M/R\beta - 1$ is a constant exponent [see Eq. (26)]. Using the isothermal model Eq. (27) above the tropopause, the pressure drops exponentially from its initial value at the tropopause P_T:

$$P = P_T \exp\left[-\frac{gM}{RT_T}(h - h_T)\right]$$

where the subscript T refers to the values at the tropopause.

The wet refraction is dependent on the partial pressure e, which decreases in somewhat the same way as total pressure in the troposphere Eq. (25) although much more rapidly. Saastamoinen[11,20] uses an exponent four times as large as in Eq. (25) to account for this difference, $e = e_0(T/T_0)^{-4gM/R\beta}$.

Saastamoinen[11,20] described both a precision model and a standard model for the tropospheric delay. Only the standard model is given here, and its delay correction for radio frequency ranging for elevation angles $E \geq 10$ deg is as follows[11,20]:

$$\Delta = 0.002277\,(1 + D)\sec\psi_0\left[P_0 + \left(\frac{1255}{T_0} + 0.005\right)e_0 - B\tan^2\psi_0\right] + \delta_R \text{ m} \qquad (30)$$

where Δ is the delay correction in meters; P_0, e_0 are in millibars; and T_0 is in °K. The correction terms B and δ_R are given in Table 4 for various user heights h. The apparent zenith angle $\psi_0 = 90$ deg $- E$. The value of D in Eq. (30) is $D = 0.0026\cos 2\phi + 0.00028h$, where ϕ is the local latitude, and h is the station height in km.

B. Hopfield Two Quartic Model

Hopfield[18,19,21] has developed a dual quartic zenith model of the refraction, different quartics for the dry and wet atmospheric profiles. Black[22,23] has extended this zenith model to add the elevation angle mapping function, as shown in a later paragraph.

Figure 8 shows an example of the measured variation in the vertical integrated index of refractivity, the zenith excess delay.[18] As can be seen, the dry component is on the order of 2.3 m and has a relatively small variation, whereas the wet

Table 4 Correction terms for Saastamoinen's[11,20] standard model, Eq. (30)

	Apparent zenith	Station height above sea level							
	Angle	0 km	0.5 km	1 km	1.5 km	2 km	3 km	4 km	5 km
	60 deg 00 min,	+0.003	+0.003	+0.002	+0.002	+0.002	+0.002	+0.001	+0.001
	66 deg 00 min,	+0.006	+0.006	+0.005	+0.005	+0.004	+0.003	+0.003	+0.002
	70 deg 00 min,	+0.012	+0.011	+0.010	+0.009	+0.008	+0.006	+0.005	+0.004
	73 deg 00 min,	+0.020	+0.018	+0.017	+0.015	+0.013	+0.011	+0.009	+0.007
	75 deg 00 min,	+0.031	+0.028	+0.025	+0.023	+0.021	+0.017	+0.014	+0.011
δ_R, m	76 deg 00 min,	+0.039	+0.035	+0.032	+0.029	+0.026	+0.021	+0.017	+0.014
	77 deg 00 min,	+0.050	+0.045	+0.041	+0.037	+0.033	+0.027	+0.022	+0.018
	78 deg 00 min,	+0.065	+0.059	+0.054	+0.049	+0.044	+0.036	+0.030	+0.024
	78 deg 30 min,	+0.075	+0.068	+0.062	+0.056	+0.051	+0.042	+0.034	+0.028
	79 deg 00 min,	+0.087	+0.079	+0.072	+0.065	+0.059	+0.049	+0.040	+0.033
	79 deg 30 min,	+0.102	+0.093	+0.085	+0.077	+0.070	+0.058	+0.047	+0.039
	79 deg 45 min,	+0.111	+0.101	+0.092	+0.083	+0.076	+0.063	+0.052	+0.043
	80 deg 00 min,	+0.121	+0.110	+0.100	+0.091	+0.083	+0.068	+0.056	+0.047
B mb		1.156	1.079	1.006	0.938	0.874	0.757	0.654	0.563

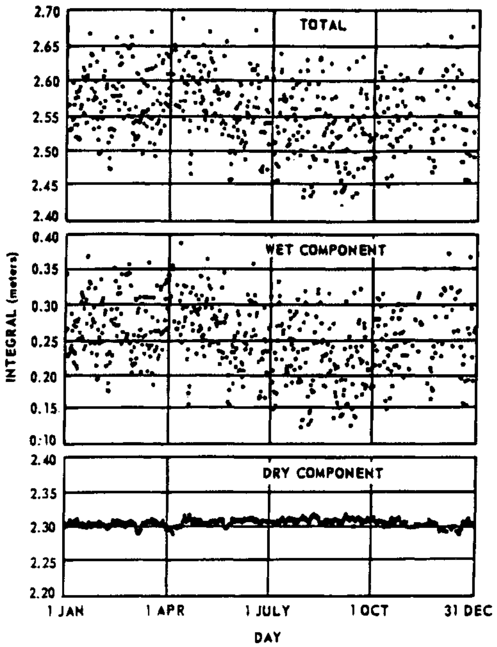


Fig. 8 Vertical integral of refractivity at Pago Pago, Samoa during 1967, balloon data from Hopfield.[18]

component is on the order of 0.25 m but varies ± 40%. Note that the scales used in all three plots are the same 5 cm/division. In data taken overland or in cold climates, the wet component is often much lower, ranging from 0 to 0.1 m.

The two-quartic model for the refractive index vs altitude h is similar in form to that of Eq. (26). Each fourth power term represents the dry and wet refractive indices as follows:

$$N_d(h) = N_{d0}\left(1 - \frac{h}{h_d}\right)^4 \quad \text{for} \quad h \le h_d = 43 \text{ km} \tag{31a}$$

and

$$N_w(h) = N_{w0}\left(1 - \frac{h}{h_w}\right)^4 \quad \text{for} \quad h \le h_w = 12 \text{ km} \tag{31b}$$

where N_{d0} and N_{w0} are the respective dry and wet refractive indexes at the surface, and h is the altitude. The zenith delay is then obtained by integrating Eq. (31) as follows:

$$\begin{aligned}\Delta &= 10^{-6}\int_0^{h_d} N_{d0}\left(1 - \frac{h}{h_d}\right)^4 \mathrm{d}h + 10^{-6}\int_0^{h_w} N_{wo}\left(1 - \frac{h}{h_w}\right)^4 \mathrm{d}h \\ &= \frac{10^{-6}}{5}[N_{d0}h_d + N_{w0}h_w] = \Delta_d + \Delta_w \end{aligned} \tag{32a}$$

Note that h_d and h_w refer to height above the surface level where the surface refractivities N_{d0} and N_{w0} are measured. This quartic relationship is taken from the constant lapse rate model in Eq. (25), which gives an exponent $-(Mg/R\beta) - 1$. As mentioned earlier, this exponent becomes 4 at a lapse rate $\beta = -6.81$°K/km. Hopfield[18,19,21] chose the 4th power for ease of calculation and as a good approximation to the observations, and integrate along the satellite to user ray path for the wet atmosphere (troposphere only) to $h_w = 12$ km and the dry atmosphere to approximately $h_d = 43$ km to cover the troposphere and stratosphere. The values of h_d and h_w are determined by best fit to experimental measurements. As shown earlier in Eq. (29), the zenith integral of dry (really hydrostatic delay) tropospheric delay is equal to a constant times the surface pressure:

$$\Delta_d = kP_0 \tag{32b}$$

where $k = (77.6\ R/g) \times 10^{-9}$ in cgs units (Ref. 19).

C. Black and Eisner (B&E) Model

Black and Eisner[23] began with a model for excess delay for paths from point P_1 to P_2 at various elevation angles, as follows:

$$\Delta = 10^{-6}\int_{P_1}^{P_2} [N_d(s) + N_w(s)]\ \mathrm{d}s \tag{33}$$

where the refraction equation is a simplified version of Eq. (17); namely, $N_d = 77.6\ P/T$ (see Eq. 32), and $N_w = 3.73 \times 10^5\ (e/T^2)$. For a straight geometric line path that neglects ray-bending effects at elevation angle E, the differential ds has been shown [simplified version of Eq. (1) for small a] to be $\mathrm{d}s = \mathrm{d}h/\sqrt{1 - [\cos E/(1 + a)]^2}$, and the excess delay becomes as follows:

$$\Delta = 10^{-6} \int_0^{h_d} \frac{(N_d(h) + N_w(h))\ \mathrm{d}h}{\sqrt{1 - [\cos E/(1 + a)]^2}} \tag{34}$$

where the h_d = 45 km; $a \overset{\Delta}{=} h/R_e$, and N_w is nonzero only for $h < h_w$ = 13 km.

Figure 9 shows the Haystack observatory measurements (in summer) of the profile of water vapor density vs height. Note the very substantial 1σ variation of the water vapor density measurements. The expression for delay error Δ, Eq. (34), is now approximated by the product of the zenith delay and a mapping function $m(E)$. The zenith delay approximation is based on numerical integration of a quartic dry term modeled after Hopfield[18,19,21] and a wet term that is exponential with saturated conditions at the surface and models measured profiles similar to Fig. 9.

The B&E delay error[23] is thus approximated by the following:

$$\Delta = (\Delta_{\mathrm{dz}} + \Delta_{wz})\ m(E, T) \tag{35}$$

where B&E use a single mapping function (as opposed to separate mapping functions for wet and dry). The mapping $m(E,T)$ is a function of both the elevation E and has the following small temperature dependence:

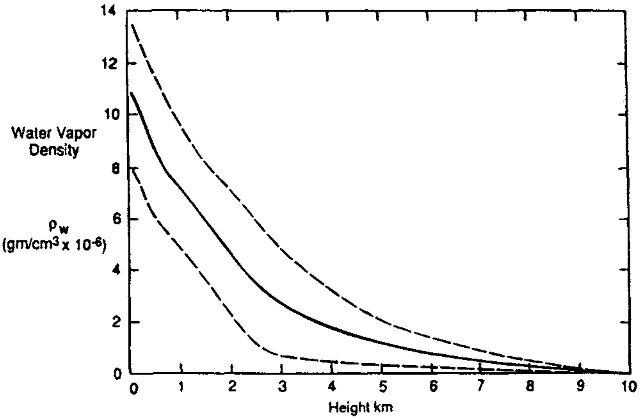


Fig. 9 Mean profile of water vapor density ρ_w taken from 45 radiosonde measurements at Haystack Observatory in August 1973. The dashed lines depict the 1 σ standard deviation. "The profile approximates an exponential dependence with a scale height of 2.2 km although the high part decreases more slowly than an exponential."[23,24]

$$m(E,\,T) = 1/\sqrt{1 - ((\cos E)/(1 + X_{dw}h_d/R_e))^2} \tag{36}$$

where X_{dw} in general varies with temperature.

For elevation angles in the range 7 deg $< E \leq 90$ deg and surface temperatures in the region $-30°\text{C} < T_0 < 40°\text{C}$, the value of X_{dw} is in the range, $0.00088 \leq X_{d_w}h_d/R_e \leq 0.001$. Thus, B&E approximate $m(E,T) \cong m(E)$

$$m(E) = 1/\sqrt{1 - ((\cos E)/(1 + 0.001))^2} = 1.001/\sqrt{(.001)^2 + 0.002 + \sin^2 E} \tag{37}$$

Compared with the results for more exact models also developed by Black and Eisner[23] this approximate model showed an error of approximately 7 cm at 7-deg elevation angle and 1 cm at 13-deg elevation angle.

D. Water Vapor Zenith Delay Model—Berman

Berman[25] has developed a simple model for water vapor zenith delay based on the correlation between the dry gas and water vapor delay for day and night measurements at Edwards Air Force Base, CA. Thus, if we can estimate the dry zenith delay from surface pressure measurements and estimate the surface wet and dry refractivities, we can then estimate the wet zenith delay as some fraction of that delay. The Berman model[25] gives the ratio of zenith excess delays for wet and dry components, $\tau_w/\tau_d = k\,N_{\omega 0}/N_{d0}$, as proportional to the ratio of wet and dry surface refractive indexes where the empirically determined scale factor is as follows:

$$k = \begin{cases} 0.2896 & \text{day} \\ 0.3773 & \text{night} \\ 0.3224 & \text{mixed day/night} \end{cases}$$

Note that this empirical result was obtained in a relatively dry region of California.

E. Davis, Chao, and Marini Mapping Functions

As discussed earlier, the mapping function attempts to relate accurately the actual excess path delay $\Delta(E)$ as a function of elevation angle and meteorological conditions to the zenith excess delay Δ_z. The mapping function can be used in conjunction with other models for the zenith delay (e.g., Davis et al.[1] and Hopfield[18]). For elevation angles near 90 deg (the zenith), clearly, a simple 1/sin E approximation is generally sufficient. However, at lower elevation angles, the nonuniform and finite width spherical shell model troposphere make this simple model inadequate.

Marini[26] described a continued fraction version of the mapping function as follows:

$$m(E) = \cfrac{1}{\sin E + \cfrac{a}{\sin E + \cfrac{b}{\sin E + \cfrac{c}{\sin E + \dots\dots\dots}}}} \tag{38}$$

where the *a, b, c.,* . . . are constants. Chao[27] employed a model of this form for the excess delay Δ that utilizes separate mapping functions for the wet and dry components. These mapping functions are of the same form as the continued fraction representation except that only the first two terms are used and the second sin *E* has been replaced by tan *E* so as to make *m*(90 deg) = 1. The specific dry and wet mapping functions used by Chao are as follows:

$$m_d(E) = \frac{1}{\sin E + \dfrac{0.00143}{\tan E + 0.0445}}$$
$$m_w(E) = \frac{1}{\sin E + \dfrac{0.00035}{\tan E + 0.017}} \tag{39}$$

Davis[1] has developed a more sophisticated function for the dry (hydrostatic) mapping function wherein the coefficients *a, b,* and *c* are dependent on surface pressures, temperature, lapse rates, and tropospheric height h_T. This representation has been compared with ray-tracing results and found to be accurate to within about 2.5 cm even at elevation angles as low as 5 deg. This model is termed the Davis Cfa 2.2 mapping function and is as follows:

$$m(E) = \frac{1}{\sin E + \dfrac{a}{\tan E + \dfrac{b}{\tan E + c}}} \tag{40}$$

where *a, b,* and *c* are dependent upon measurements or estimates[1]:

$$\begin{aligned} a = 0.001185[1 &+ 0.6071 \times 10^{-4}(P_0 - 1000) - 0.1471 \times 10^{-3}\, e_0 \\ &+ 0.3072 \times 10^{-2}(T_0 - 20) + 0.1965 \times 10^{-1}(\beta + 6.5) \\ &- 0.5645 \times 10^{-2}(h_T - 11.231)] \\ b = 0.001144[1 &+ 0.1164 \times 10^{-4}(P_0 - 1000) - 0.2795 \times 10^{-3}\, e_0 \\ &+ 0.3109 \times 10^{-2}(T_0 - 20) + 0.3038 \times 10^{-1}(\beta + 6.5) \\ &- 0.1217 \times 10^{-1}(h_T - 11.231)] \end{aligned} \tag{40a}$$

$$c = -0.0090$$

and β is the tropospheric temperature lapse rate in °C/km; h_T is the height of the tropopause in km; T_0 is the surface temperature in °C; and P_0 is the surface pressure in mb.

F. Altshuler and Kalaghan Delay Model

This tropospheric excess delay model[28] is easy to use and that can either use monthly average refractivity values or can be updated with real-time corrections. The model, hereafter called the A&K model (note that there is an errata sheet in the published report for this model), consists of empirical expressions for tropospheric range error corrections suitable for GPS users, based upon worldwide statistics of surface refractivity. The model requires that the user input the elevation angle, height above the surface, and the surface refractivity, if known. If the user cannot measure the actual surface refractivity, the A&K model has an expression for the average surface refractivity as a function of height above sea level, latitude, and season of the year. The A&K model also contains a term for off-zenith angle correction, valid down to a viewing angle of 3-deg elevation angle. The A&K model has the following form:

$$\begin{aligned}\Delta(E, h, N_s) &= 2.29286\, m(E)H(h)F(h, N_s) \qquad \text{feet}\\ &= (0.3048)(2.29286)\, m(E)H(h)F(h, N_s) \qquad \text{meter} \qquad (41)\end{aligned}$$

where $m(E)$ is the A&K mapping function, and H is a function of the user's height (note that one foot is exactly 0.3048m). The function F depends on both the user's height and the surface refractivity N_s. The factors are chosen to be of polynomial form to simplify computation:

$$\begin{aligned}m(E) = \{&(0.1556 + 138.8926/E - 105.0574/E^2 + 31.5070/E^3)\\ &+ [1.000 + 1.0 \times 10^{-4}(E - 30.00)^2]\}/2.29286 \qquad (41a)\end{aligned}$$

where E is in deg. [The factor 2.29286 has been broken out from the mapping function in the original A&K formulation, so that $m(90 \text{ deg}) = 1$.] The function $H(h)$ is as follows:

$$\begin{aligned}H(h) = [&(0.00970 - 2.08809(h + 8.6286)^{-1} + 122.73592(h + 8.6286)^{-2}\\ &- 703.82166(h + 8.6286)^{-3}] \qquad (41b)\end{aligned}$$

where h is in thousands of feet. The third function $F(h,N_s)$ is as follows:

$$\begin{aligned}F(h, N_s) = 3.28084&\left[\frac{6.81758}{h + 3.28084} + 0.30480(h + 3.28084) + 0.00423\, N_s\right.\\ &\left. - 1.33333\right][1 - 1.41723 \times 10^{-6}(N_s - 315.000)] \qquad (41c)\end{aligned}$$

The surface refractivity N_s, if not known, can be estimated as follows:

$$N_s = \alpha_0 + \alpha_1 h' + \alpha_2\phi + \alpha_3 h's^2 + \alpha_4\phi s^2 + \alpha_5 h'c + \alpha_6\phi\lambda c \qquad (41d)$$

where ϕ is the latitude in degrees, and h' is the height of the surface above sea level in feet. The constants α_i and c are as follows: $\alpha_0 = 369.03$; $\alpha_1 = -0.01553$; $\alpha_2 = -0.92442$; $\alpha_3 = -0.0016$; $\alpha_4 = -0.19361$; $\alpha_5 = 0.00063$; $\alpha_6 = -0.05958$; $s = \sin(\pi M/12)$; and $c = \cos(\pi M/12)$. The value M (month) varies with the season: $M = 1.5$ winter; 4.5_spring; 7.5 summer; and 10.5 fall. The average global surface refractivity[28] is $\overline{N} = 324.8$, and the standard deviation is $\sigma_N = 30.1$.

Altshuler and Kalaghan[28] have estimated errors in their tropospheric range error model to be approximately 3.7% if the surface refractivity is known, and 6% if it is not known. If only a global average surface refractivity of 324.8 N units is used, they estimate the standard error of their model to be approximately 8%.

G. Ray Tracing and Simplified Models

For some purposes, a much simpler model of the tropospheric delay is all that is required. Typical troposphere delays vs elevation angle for a user at sea level have been computed by Janes et al.[7] by ray tracing for U.S. Standard Atmosphere and are shown in Table 5. Notice that the total delay at zenith ranges from a low of 2.316 m at latitude 75°N in January to a high of 2.576 m at 30°N in July. Shown in the lower portion of Table 5, the dry atmosphere delay ranged from 2.297 m in July at 60°N to a high of 2.328 m in January at 30°N for a 3-cm variation. The wet delay at zenith varied from a low of 0.015 m in January at 75°N to a high of 0.263 m in July at 30°N for a 24.8-cm variation. Notice that the ratio of 5-deg elevation angle delay to 90-deg elevation angle delay is fairly constant ranging from 10.183 to 10.231 at 75°N but that at the midlatitude of 30–45°, the ratio ranges only from 10.187 to 10.201 for only a 0.14% variation. Note the ratio $\Delta_w(90^\circ)$ to $\Delta(90^\circ)$ of wet delay to total delay is approximately 7% in summer in the midlatitudes (45–75°) but is only 0.6–2% in the winter.

If a simple model for sea level delay is employed, the excess delay could be modeled as follows:

$$\Delta = 2.47/(\sin E + 0.0121) \tag{42}$$

which gives a value at zenith of 2.44 that is the average value for users in Table 3 at 30°N, 45°N, 60°N, and gives a mapping function value of approximately 10.2 at a 5-deg elevation angle.

Figure 10 shows this simplified mapping function

$$m(E) = 1.0121/(\sin E + 0.0121)$$

and the mapping functions of Black and Eisner[23]

$$m(E) = 1/\sqrt{1 - (\cos E/1.001)^2},$$

and Altshuler and Kalaghan[28]. They are relatively close to one another on the graph.

If we assume that refractivity varies as $(1 - h/h_d)^4 = (1 - .023\mathrm{h})^4$ for h in km and a scale height $h_d = 43$ km, then the zenith delay for a user at altitude h is computed as follows:

$$\Delta(h) = N_s \int_h^{h_d} (1 - h/h_d)^4 \, \mathrm{d}h = \Delta(0)(1 - h/h_d)^5 \tag{43}$$

As an alternative approach, an exponential profile can be employed to give the following simplified model:

$$\Delta(h) = 2.47\, e^{-0.133h}/(\sin E + 0.0121) \tag{44}$$

so that the zenith delay decreases to $e^{-0.133h} = 0.393$ of the sea level value at $h = 7$ km altitude.

H. Lanyi Mapping Function and GPS Control Segment Delay Estimate

Lanyi[29,30] has published one of the more precise but complex estimates of the tropospheric delay vs elevation angle where excess delay is of the form

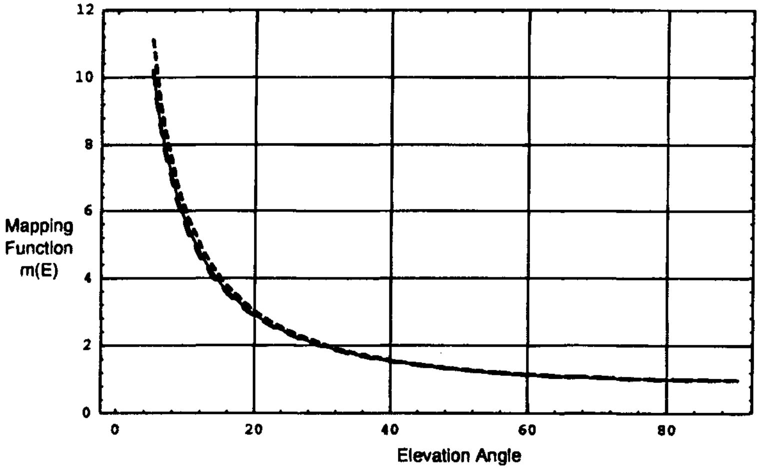


Fig. 10 Mapping functions of the simplified model (dashed), the B&E[23] (fine solid line) model, and the A&K[28] model (fine dashed line) for the range of E $5 < E \leq 90$ deg. The values of m(5 deg) are 10.1969, 10.218, and 11.114 for the simplified model, Eq. (44) the B&E model, and the A&K model, respectively.

$$\Delta(E) = F(E)/\sin E \qquad E > 5^\circ$$

where

$$F(E) = \Delta_d F_d(E) + \Delta_w F_w(E) + (\Delta_d^2/D)F_{b1}(E) + (2\Delta_d\Delta_w/D)F_{b2}(E) + \Delta_w^2/D)F_{b3}(E) + (\Delta_d^3/\Delta^2)F_b4(E).$$

The terms Δ_w and Δ_d are the wet and dry zenith delays, respectively, D is a scale height ≈8.567 km, and the F_{b1}, F_{b2}, F_{b3}, F_{b4}, are various complicated mapping functions. Sufficė it to say that special cases of the Lanyi mapping function yield approximations to the Saastamoinen and Black mapping functions.

The GPS Control Segment also employs a model for tropospheric excess delay to correct the smoothed pseudoranges. The overall equation for tropospheric delay is[31]:

$$\Delta(E) = \frac{(0.02312)P_s[\mathrm{T} - 4.11 + 5(r_m - r_a)/148.98]}{T\sqrt{1 - \left[\dfrac{r_a \cos E}{r_m + (1 - C)(r_d - r_m)}\right]^2}} + \frac{(0.0746)eh[1 + 5(r_m - r_a/h]}{T^2\sqrt{1 - \left[\dfrac{r_a \cos E}{r_m + (1 - C)h}\right]^2}} \qquad E > 5^\circ$$

where P_s = monitor station barometric pressure converted to kilopascals, T = monitor station measured temperature converted to Kelvin, r_m = radial distance

Table 5 Troposphere delay (wet + dry component) in meters vs elevation angle in degrees, latitude, and season developed by ray tracing the U.S. Standard Atmosphere; also shown in the lower portion of the table is the individual wet and dry delay components (using results from Janes et al.[7])

Elevation angle E	15°N annual avg.	30°N		45°N		60°N		75°N	
		July	Jan.	July	Jan.	July	Jan.	July	Jan.
5 deg	26.088	26.279	25.06	25.365	24.239	24.781	23.853	24.409	23.694
10 deg	14.237	14.341	13.689	13.854	13.234	13.352	13.002	13.338	12.911
20 deg	7.416	7.47	7.132	7.219	6.893	7.05	6.769	6.95	6.719
90 deg = zenith	2.558	2.576	2.46	2.49	2.378	2.431	2.334	2.397	2.316
Ratio, Δ(5 deg)/ Δ(90 deg)	10.199	10.201	10.187	10.187	10.193	10.194	10.220	10.183	10.231
Ratio, Δ_w(90 deg)/ Δ(90 deg)	0.0958	0.1020	0.0537	0.0731	0.0240	0.0551	0.0124	0.0405	0.0065

Elevation angle	Dry/wet excess delay, m								
		30°N latitude		45°N latitude		60°N latitude		75°N latitude	
	Annual	July	Jan.	July	Jan.	July	Jan.	July	Jan.
90 deg	2.313/.245	2.313/.263	2.328/.132	2.308/.182	2.321/.057	2.297/.134	2.305/.029	2.300/.097	2.301/.015
20 deg	6.702/.714	6.703/.767	6.747/.385	6.688/.531	6.727/.166	6.658/.392	6.683/.086	6.667/.283	6.674/.045
10 deg	12.841/1.396	12.842/1.499	12.936/.753	12.817/1.037	12.910/.324	12.766/.766	12.835/.167	12.786/.552	12.823/.088
5 deg	23.381/2.707	23.379/2.900	23.600/1.460	23.356/2.009	23.612/.627	23.300/1.481	23.530/.323	23.341/1.068	23.523/.171

from the Earth's center to the meterorological sensors, r_a = radial distance from the Earth's center to the monitor station antenna, r_d = tropospheric dry radius, E = satellite elevation, h = wet height of troposphere, e = estimate of the partial water vapor pressure, C = assumed integration constant = 0.85 when $E > 5.0°$.[31]

I. Model Comparisons

Janes et al.[7] have done an extensive comparison of the various models described in the previous paragraphs. A few words of summary are appropriate here. For the zenith delay, both the Saastamoinen and Hopfield models in Refs. 11 and 18–21 gave results for the dry component that were within several mm of the ray trace delay as computed for the U.S. Standard Atmosphere (Table 5). Even at a 5-deg elevation angle, the Hopfield model was within 5 cm, whereas the two-layer model[11,20] gave more accurate results to within 6 mm. At zenith, the water vapor component for the standard model[11,20] gave results mostly within 30 mm, whereas the model errors[18,19,21] were within 20 mm.

Janes also compared the mapping function error $\delta m(E)$ scaled by the total zenith delay; namely, $\delta m(E)\Delta_z$. For the total delay at the 20-deg elevation angle, the Saastamoinen and Black and Eisner models of Refs. 11, 20, and 23 gave results within 8 mm. These errors rose to approximately 50 mm at 10 deg elevation. At 5-deg elevation angle the result[23] was still within 10 cm, whereas the standard model[11,20] produced errors of 1.2 m. For the dry gas component the Davis mapping function was the most accurate, yielding errors of less than 6 cm at 5-deg elevation angle.

J. Tropospheric Delay Errors and GPS Positioning

Errors in the estimates of the tropospheric delay have some amount of cancellation relative to GPS positioning because of the correlation in the errors for different satellites. Clearly, a constant bias for four pseudoranges to four different satellites would not affect position errors at all but would cause a user clock time bias error. As we have seen, the significant tropospheric errors are likely to be those from satellites at low elevation angles. If, as an example, four satellites at low elevation angle all have an identical error bias in the troposphere delay, and a satellite at zenith has no delay bias error, the bias error in the horizontal plane position estimate cancels out, but an altitude error remains.

One of the primary contributors to horizontal position error, in addition to lower elevation angle prediction of the dry tropospheric delay (with satellites of different elevation angles) is the variation in wet low elevation angle errors for satellites at substantially different azimuths. At 5-deg elevation angle, the wet troposphere can contribute widely varying delay errors (relative to the total wet delay contribution ≈ 25 cm). The water vapor zenith correction can also change significantly 2 cm with a 10-km horizontal displacement, or in a few hours of time passage, and these can be magnified by the mapping function at low elevation angles E. Water vapor zenith delay changes of 3 cm have been observed in one hour.[3] There is also no way to predict the wet delay correction change with the azimuth. It is expected, however, that the dry component of delay error would be highly correlated with azimuth; only the wet component is expected to decorrelate in any significant way with azimuth.

References

[1]Davis, J. L., et al., "Geodesy by Radio Interferometry: Effects of Atmospheric Modeling Errors on Estimates of Baseline Length," *Radio Science,* Vol. 20, 1985.

[2]NOAA, "U.S. Standard Atmosphere, 1976," NOAA-S/T76-1562, National Technical Information Service, U.S. Dept. of Commerce, Springfield, VA, 1976.

[3]Elgered, G., "Tropospheric Radio Path Delay," edited by M. A. Jansse, *Atmospheric Remote Sensing by Microwave Radiometry,* Wiley, New York, 1993.

[4]Smith, D. G., (ed.), The Cambridge Encyclopedic of Earth Sciences, Cambridge University Press, New York, 1981.

[5]Boithias, L., *Radio Wave Propagation,* McGraw Hill, New York, 1982.

[6]Spilker, J. J., Jr., "Digital Communications by Satellite," Prentice Hall, Englewood Cliffs, NJ, 1977, 1995.

[7]Janes, H. W., Langley, R. B., and Newby, S. P., "Analysis of Tropospheric Delay Prediction Models," *Bulletin Géodésiqué,* Vol. 65, 1991.

[8]Ippolito, L. J., Jr., *Radio Wave Propagation in Satellite Communications,* Van Nostrand, New York, 1986.

[9]Moulsley, T. J., and Vilar, E. "Experimental and Theoretical Statistics of Microwave Amplitude Scintillation on Satellite Downlinks," *IEEE Trans. Antennas and Propagation,* Vol. AP-30, pp. 1099–1106, Nov. 1982.

[10]Sears, F. W., *Optics,* Addison-Wesley, Reading, MA, 1949.

[11]Saastamoinen, J., "Contribution to the Theory of Atmospheric Refraction, *Bulletin Géodésiqué,* Vol. 105, Sept. 1972, Vol. 106, Dec. 1972, Vol. 107, March 1973.

[12]Thayer, G. D., "An Improved Equation for the Radio Refractive Index of Air," *Radio Science,* Vol. 9, 1974.

[13]Hill, R. J., Lawrence, R. S., and Priestly, J. T., "Theoretical and Calculated Aspects of the Radio Refractive Index of Water Vapor," *Radio Science,* Vol. 17, pp. 1251–1257, 1982.

[14]Birmbaum, G., and Chatterjee, S. K., "The Dielectric Constant of Water Vapor in the Microwave Region," *Journal of Applied Physics,* Vol. 23, 1952.

[15]Boudouris, G., "On the Index of Retraction of Air, the Absorption and Dispersion of Centimeter Waves by Gases," *Journal of Research of the National Bureau of Standards,* 1963.

[16]Sears, F. W., *Thermodynamics, The Kinetic Theory of Gases, and Statistical Mechanics,* Addison-Wesley, Reading, MA, 1953.

[17]Abbott, M. M., and Van Ness, H. C., *Thermodynamics,* McGraw Hill, New York, 1989.

[18]Hopfield, H. S., "Tropospheric Effect on Electromagnetically Measured Range: Prediction from Surface Weather Data," Applied Physics Laboratory, Johns Hopkins University, Baltimore, MD, July 1970.

[19]Hopfield, H. S., "Tropospheric Range Error Parameters—Further Studies," Applied Physics Laboratory, Johns Hopkins University, Baltimore, MD, June 1972.

[20]Saastamoinen, J., "Atmospheric Correction for the Troposphere and Stratosphere in Radio Ranging of Satellites," Geophysical Monograph 15, American Geophysical Union, Washington, DC, 1972.

[21]Hopfield, H. S., "Two Quartic Tropospheric Refractivity Profile for Correcting Satellite Data," *Journal of Geophysical Research,* April 1969.

[22]Black, H. D., "An Easily Implemented Algorithm for the Tropospheric Range Correction," *Journal of Geophysical Research,* Vol. 38, No. 4, 1978, pp. 1825–1828.

[23] Black, H. D., and Eisner A., "Correcting Satellite Doppler Data for Tropospheric Effects," *Journal of Geophysical Research,* Vol. 89, 1984.

[24] Moran, J. M., and Rosen, B. R., "The Estimation of the Propagation Delay Through the Troposphere from Microwave Radiometer Data," *Radio Interferometry Techniques for Geodesy,* NASA Conf. Publ. 2125, 1979.

[25] Berman, A. L., "The Prediction of Zenith Refraction from Surface Measurements of Meteorological Parameters," JPL TR-32-1602 California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, 1976.

[26] Marini, J. W., "Correction of Satellite Tracking Data for an Arbitrary Tropospheric Profile," *Radio Science,* Vol. 7, 1972.

[27] Chao, C. C., "The Tropospheric Calibration Model for Mariner Mars, 1971," JPL TR 32-1587, Jet Propulsion Laboratory, Pasadena, CA, March 1974.

[28] Altshuler, E. E., and Kalaghan, P. M., "Tropospheric Range Error Corrections for the NAVSTAR System," Microwave Physics Laboratory, Air Force Cambridge Research Laboratories, April 1974.

[29] Lanyi, G., "Tropospheric Delay Effects in Radio Interferometry," Telecommunications and Data Acquisition Progress Rep., Jet Propulsion Lab, Pasadena, CA, April–June 1984.

[30] Sovers, O. J., and Lanyi, G. E., "Evaluation of Current Tropospheric Mapping Functions by Deep Space Network Very Long Baseline Interferometry," The Telecommunications and Data Acquisition Progress Rep. 42–119, Jet Propulsion Laboratory, Pasadena, CA, Nov. 1994.

[31] Master Control Station Kalman Filter–Mission Support Study Guide, 2nd ed. Oct. 1993.

Bibliography

Brunner, F. K., ed., "Atmospheric Effects on Geodetic Space Measurements," Monograph 12, School of Surveying, University of New South Wales, Kinsington, NSW, Australia, 1988.

Hendy, M. R., and Brunner, F. K., "Modeling the Zenith Wet Component of the Tropospheric Path Delay for Microwaves," *Australian Journal of Geodesy, Photogrammetry & Surveying,* Dec. 1990.

Chapter 14

Multipath Effects

Michael S. Braasch*
Ohio University, Athens, Ohio 45701

I. Introduction

THE Global Positioning System (GPS) has been shown to be capable of supporting a wide variety of exciting applications. In addition to the usual functions of position, velocity, and time determination, it is also possible to perform attitude and heading determination of dynamic platforms, measurement of flexing in large space structures, and to provide precision approach and landing guidance. The later applications require high accuracy in real time. However, these technologies face a major stumbling block. It is the effect of multipath. Multipath is the phenomenon whereby a signal arrives at a receiver via multiple paths attributable to reflection and diffraction. Multipath represents the dominant error source in satellite-based precision guidance systems.

Multipath has been cited as a major error source both in differential satellite systems as well as interferometry.[1,2] Multipath distorts the signal modulation and degrades accuracy in conventional and differential systems. Multipath also distorts the phase of the carrier, and hence degrades the accuracy of the interferometric systems. Furthermore, because interferometric systems often employ pseudorange measurements for initialization (ambiguity resolution) purposes, multipath contamination of the pseudorange can increase the time required for initialization.

For standard code-based differential systems, signal degradation attributable to multipath can be severe. This stems from the fact that multipath is a highly localized phenomenon. Multipath sources that affect the ground reference station receiver do not necessarily cause errors in the mobile receiver. Likewise, multipath sources that affect the mobile receiver do not necessarily affect the ground reference station.

Multipath effects in pseudorandom noise (PRN) ranging have been studied since the early 1970s. Hagerman[3] derived relationships involving multipath and PRN code-tracking error. This fundamental work formed the basis for the analysis of GPS code and carrier multipath errors during the field tests at the Yuma Proving Ground (YPG).[4] In the early 1980s, the effect of multipath on short

* Assistant Professor, Department of Electrical and Computer Engineering.

baseline interferometry was studied at the Massachusetts Institute of Technology[5,6] and at the Charles Stark Draper Laboratory.[7] The studies concluded that the effects of multipath (in this differential carrier-phase tracking system) could be reduced to a few centimeters of error over short baselines if the signals could be averaged over a period of an hour or more. Bletzacker[8] also considered multipath errors in geodetic applications. Performance improvements were obtained by mounting the antenna on rf-absorbing material, thereby improving the characteristics of the antenna pattern. Tranquilla and Carr[9] confirmed this by collecting data in stressful environments using a geodetic antenna with and without an rf-absorbing ground plane.

Falkenberg et al.[10] and Lachapelle et al.[1] describe marine differential GPS experiments in which multipath was mitigated through the use of rf-absorbing ground planes and filtering schemes. Evans[11] demonstrated multipath effects on ionospherically corrected code and carrier measurements from a geodetic GPS receiver. Georgiadou and Kleusberg[12] considered multiple reflections and showed that multipath on short baselines could be detected using dual-frequency measurements. Abidin[13] examined the effects of multipath in dual-frequency-measurement-based ambiguity resolution.

The effect of multipath on ionospheric measurements using GPS was presented by Bishop et al.[14] Their work verified the theoretical multipath relations derived by Hagerman[3] and considered various mitigation schemes for static applications. Sennott and Pietraszewski[15] and Sennot and Spalding[16] have developed state variable models for the estimation and mitigation of multipath in differential GPS ground reference stations.

Van Nee[17] has shown that code-phase multipath error traces tend not to be zero mean and can have periods on the order of an hour. This contradicts the popular notion that code-phase multipath can be eliminated in static applications simply through averaging. Van Nee[18,19] has shown that an exception to this rule occurs when a coherent delay–lock-loop (DLL) receiver is used and the rate-of-change of the multipath relative phase is large compared to the tracking loop bandwidth.

Much of this chapter is based upon Braasch,[20] which examined the characteristics of multipath error in the precision approach and landing environment.

II. Signal and Multipath Error Models

Fundamental to the understanding of the effects of multipath in any given environment is an understanding of PRN ranging receivers and how multipath distortion results in ranging errors. This section derives closed form expressions for code-phase and carrier-phase multipath errors resulting from a single multipath ray entering a stationary receiver. Although most multipath scenarios involve multiple rays, much insight results from the analysis of the single-ray case. It also serves as the starting point from which the multiple-ray case can be considered. The analysis considers the two most prevalent types of receivers: the coherent delay–lock-loop (DLL) and the noncoherent DLL. This section extends the theoretical developments documented by Hagerman[3] to include consideration of code multipath error spectra. It should be noted that although the derivations follow the signal flow of analog receivers, the results are exactly the same for modern digital architectures.

A. Pseudorandom Noise Modulated Signal Description

The signal broadcast from the satellite in a PRN ranging system may be expressed as follows:

$$s_1(t) = A\cos[\omega_o t + p(t)\pi/2] \tag{1}$$

where $0.5*A^2$ is the average signal power into the receiver, ω_o is the frequency of the received signal in radians per second (carrier frequency plus Doppler shift), and $P(t)$ is the PRN code (either $+1$ or -1).

Note that the actual GPS signal is considerably more complicated. The GPS carrier is modulated by two PRN codes (the coarse/acquisition (C/A) code and the precision (P) code) in addition to navigation data. For the present purpose of multipath analysis, however, the model given by Eq. (1) is adequate. By applying trigonometric identities, Eq. (1) may be rewritten as follows:

$$\hat{s}_1(t) = -Ap(t)\sin(\omega_o t) \tag{2}$$

Once inside the receiver, multipath is characterized by four parameters (all of which are relative to the direct signal): 1) amplitude; 2) time delay; 3) phase; 4) phase rate of change. For the present discussion, a stable multipath scenario is assumed; thus, the relative phase rate of change is assumed to be zero. Relative phase of the multipath is a function of the relative time delay and the reflection coefficient of the reflecting object. If the received signal is composed of the direct signal plus a single multipath ray, it may be expressed as follows:

$$s_{1m}(t) = \hat{s}_1(t) + \alpha\hat{s}_1(t + \delta) \tag{3}$$

where α is the multipath relative amplitude and δ is the multipath relative time delay (note: must be negative given the convention used in the equation).

Note that the relative phase of the multipath is not shown. Substitution of Eq. (2) into Eq. (3) and inclusion of relative phase yields the following:

$$s_{1m}(t) = -Ap(t)\sin(\omega_o t) - \alpha Ap(t + \delta)\sin(\omega_o t + \theta_m) \tag{4}$$

where θ_m is the multipath relative phase.

B. Coherent Pseudorandom Noise Receiver

This section derives expressions for the coherent DLL discriminator curve in the absence and presence of multipath. Also derived is the expression for the composite phase of the multipath corrupted signal as it is tracked by the carrier-tracking loop.

1. Coherent Delay Lock Loop Discriminator Curve in the Absence of Multipath

Following the usual signal flow in an analog receiver, the incoming signal $S_1(t)$ is mixed with early and late versions of the PRN code modulated onto a local oscillator frequency:

$$s_E(t) = -p(t + \tau - \tau_d)\sin(\omega_l t + \theta) \tag{5}$$

$$s_L(t) = -p(t + \tau + \tau_d)\sin(\omega_l t + \theta) \tag{6}$$

where τ is the DLL tracking error, τ_d is the time advance of the early code or the time delay of the late code (relative to the on-time code), θ is the tracking error of the phase lock loop (PLL), and ω_l is the local oscillator frequency in rad/s.

The signals output from the mixers are passed through bandpass filters (BPF). It is assumed that the passband of the BPFs is narrow enough to reject the sum-frequency terms. The filters also serve to complete the correlation process. The bars shown over the quantities in Eqs. (7) and (8) denote the operation of the BPFs. The output of the filters is as follows:

$$\overline{s_1(t)s_E(t)} = 1/2\,AR(\tau - \tau_d)\cos(\omega_1 t - \theta) \tag{7}$$

$$\overline{s_1(t)s_L(t)} = 1/2\,AR(\tau + \tau_d)\cos(\omega_1 t - \theta) \tag{8}$$

where ω_1 is the difference-frequency ($\omega_1 = \omega_o - \omega_l$) and $R(\tau)$ is the correlation function of the PRN code.

A sufficient approximation of the PRN code correlation function is given by the following:

$$\begin{aligned} R(\tau) &= 1 - \frac{|\tau|}{T}, \qquad |\tau| \le T \\ &= 0, \qquad\qquad\quad |\tau| > T \end{aligned} \tag{9}$$

where T is the PRN code bit period (note: a PRN code bit is also known as a "chip." Accordingly, the PRN code bit-rate is also known as the "chipping-rate").

Correlation sidelobes are ignored, and infinite bandwidth is assumed. The finite bandwidth of the BPFs distorts the shape of the PRN code bits and results in a smoothing of the correlation function.[23] Comparison of the results of this analysis (Sec. II. B.) with those obtained when smoothing the correlation function[24] reveals that the smoothing slightly reduces the maximum range error due to a given set of multipath parameters. As a result, Eq. (9) yields slightly conservative results.

The signals out of the BPFs are thus the early and late correlator functions at the intermediate frequency (IF). Low-pass filters can be used to remove the IF term leaving the baseband signal. Assuming a PLL carrier-phase tracking error of zero ($\theta = 0$) then yields the following:

$$\overline{s_1(t)s_E(t)} = 1/2\,AR(\tau - \tau_d) \tag{10}$$

$$\overline{s_1(t)s_L(t)} = 1/2\,AR(\tau + \tau_d) \tag{11}$$

Finally, the discriminator function is formed, for example, by differencing the outputs of the LPFs. The normalized form of the discriminator function is as follows:

$$D_c(\tau) = R(\tau + \tau_d) - R(\tau - \tau_d) \tag{12}$$

where the subscript c denotes coherent DLL operation.

Note that the DLL tracks the peak of the correlation function by tracking the zero-crossing of the discriminator function since both occur for $\tau = 0$. As derived in the next section, multipath distorts the discriminator curve so that the zero-crossing occurs for some non-zero τ. Thus, the τ corresponding to the zero-crossing is the DLL tracking error caused by the multipath. Note that although τ is an error in correlation, ranging is a time-domain function. An error in correlation corresponds to an equal but opposite error in timing. Thus, the ranging error is given by the opposite of the DLL tracking error.

2. *Coherent Delay Lock Loop Discriminator Curve in the Presence of Multipath*

In the presence of multipath, the incoming signal is given by Eq. (4). In the case of the coherent DLL, the early and late codes are modulated onto a local oscillator frequency, which is phase-locked to the incoming signal. However, the incoming signal has been perturbed because of the presence of the multipath. Thus, the PLL tracks the phase of the composite signal and not that of the direct signal. The early and late signals are now given by the following:

$$s_{Em}(t) = -p(t + \tau - \tau_d)\sin(\omega_I t + \theta_c + \theta) \tag{13}$$

$$s_{Lm}(t) = -p(t + \tau + \tau_d)\sin(\omega_I t + \theta_c + \theta) \tag{14}$$

where θ_c is the composite phase of the direct plus multipath signal and θ is the PLL carrier-phase tracking error of the composite signal.

The relations governing the composite phase of the direct plus multipath signal are derived in the next section. The outputs of the mixers are given by multiplying Eq. (4) by Eqs. (13) and (14).

Passing the output of the mixers through the BPFs completes the correlation process and removes the sum frequency terms. Low-pass filters then allow only the baseband signals to pass, yielding the following:

$$\begin{aligned}\overline{s_{1m}(t)s_{Em}(t)} &= 1/2\, AR(\tau - \tau_d)\cos(-\theta_c - \theta)\\ &\quad + 1/2\, \alpha AR(\tau - \tau_d - \delta)\cos(\theta_m - \theta_c - \theta)\end{aligned} \tag{15}$$

$$\begin{aligned}\overline{s_{1m}(t)s_{Lm}(t)} &= 1/2\, AR(\tau + \tau_d)\cos(-\theta_c - \theta)\\ &\quad + 1/2\, \alpha AR(\tau + \tau_d - \delta)\cos(\theta_m - \theta_c - \theta)\end{aligned} \tag{16}$$

Assuming a PLL tracking error of zero and taking advantage of the fact that the cosine is an even function yields the following:

$$\begin{aligned}\overline{s_{1m}(t)s_{Em}(t)} &= 1/2\, AR(\tau - \tau_d)\cos(\theta_c)\\ &\quad + 1/2\, \alpha AR(\tau - \tau_d - \delta)\cos(\theta_m - \theta_c)\end{aligned} \tag{17}$$

$$\begin{aligned}\overline{s_{1m}(t)s_{Lm}(t)} &= 1/2\, AR(\tau + \tau_d)\cos(\theta_c)\\ &\quad + 1/2\, \alpha AR(\tau + \tau_d - \delta)\cos(\theta_m - \theta_c)\end{aligned} \tag{18}$$

Again, the discriminator function is formed by differencing the outputs of the LPFs. The normalized form of the discriminator function is as follows:

$$\begin{aligned} D_{cm}(\tau) &= R(\tau + \tau_d)\cos(\theta_c) + \alpha R(\tau + \tau_d - \delta)\cos(\theta_m - \theta_c) \\ &\quad - R(\tau - \tau_d)\cos(\theta_c) - \alpha R(\tau - \tau_d - \delta)\cos(\theta_m - \theta_c) \\ &= [R(\tau + \tau_d) - R(\tau - \tau_d)]\cos(\theta_c) \\ &\quad + \alpha[R(\tau + \tau_d - \delta) - R(\tau - \tau_d - \delta)]\cos(\theta_m - \theta_c) \end{aligned} \tag{19}$$

where the subscript *cm* denotes coherent DLL operation in the presence of multipath.

3. *Carrier Phase Lock Loop Operation in the Presence of Multipath*

Prior to entering the PLL, the incoming signal is mixed with the on-time code modulated onto the local oscillator frequency. The incoming signal in the presence of a single multipath ray was given by Eq. (4). The on-time signal generated by the receiver is given by the following.

$$s_{Om}(t) = -p(t + \tau)\sin(\omega_l t + \theta_c + \theta) \tag{20}$$

After mixing the two signals and passing through a BPF, we achieve the following;

$$\begin{aligned} \overline{s_{1m}(t)s_{Om}(t)} &= 1/2\, AR(\tau)\cos((\omega_o - \omega_l)t - \theta_c - \theta) \\ &\quad + 1/2\, \alpha AR(\tau - \delta)\cos((\omega_o - \omega_l)t + \theta_m - \theta_c - \theta) \end{aligned} \tag{21}$$

Assuming perfect carrier PLL tracking of the composite signal ($\theta = 0$) and substituting $\omega_1 = \omega_o - \omega_l$, Eq. (21) becomes the following:

$$\begin{aligned} \overline{s_{1m}(t)s_{Om}(t)} &= 1/2\, AR(\tau)\cos(\omega_1 t - \theta_c) \\ &\quad + 1/2\, \alpha AR(\tau - \delta)\cos(\omega_1 t + \theta_m - \theta_c) \end{aligned} \tag{22}$$

Using the trigonometric identity for the cosine of a sum gives the following:

$$\begin{aligned} \overline{s_{1m}(t)s_{Om}(t)} &= 1/2\, AR(\tau)[\cos(\omega_1 t)\cos(\theta_c) + \sin(\omega_1 t)\sin(\theta_c)] \\ &\quad + 1/2\, \alpha AR(\tau - \delta)[\cos(\omega_1 t)\cos(\theta_c - \theta_m) + \sin(\omega_1 t)\sin(\theta_c - \theta_m)] \end{aligned} \tag{23}$$

Grouping the cosine and sine terms leads to the following:

$$\begin{aligned} \overline{s_{1m}(t)s_{Om}(t)} &= 1/2\, A[R(\tau)\cos(\theta_c) \\ &\quad + \alpha R(\tau - \delta)\cos(\theta_c - \theta_m)]\cos(\omega_1 t) + 1/2\, A[R(\tau)\sin(\theta_c) \\ &\quad + \alpha R(\tau - \delta)\sin(\theta_c - \theta_m)]\sin(\omega_1 t) \end{aligned} \tag{24}$$

The PLL tracks the composite signal so that the coefficient of the $\sin(\omega_1 t)$ term is nulled:

$$R(\tau)\sin(\theta_c) + \alpha R(\tau - \delta)\sin(\theta_c - \theta_m) = 0 \tag{25}$$

a. Region I: Absolute value of the DLL tracking error less than one chip. In this region, the correlation function $R(\tau)$ is nonzero. After making trigonometric substitutions and rearranging, Eq. (25) may be rewritten as follows:

$$\begin{aligned}\tan(\theta_c) &= \sin(\theta_c)/\cos(\theta_c)\\ &= \alpha R(\tau-\delta)\sin(\theta_m)/[R(\tau)+\alpha R(\tau-\delta)\cos(\theta_m)]\end{aligned}\tag{26}$$

For region I operation, then, the composite phase of the signal entering the PLL is given by the arctangent of the right-hand side of Eq. (26). The interdependency of the code- and carrier-tracking loops may also be observed in Eq. (26) by noting the presence of the code-correlation function.

Assuming that the multipath strength is always less than or equal to the direct, it can be shown that the carrier-phase error can be no more than 90 deg. At the GPS L_1 band (1575.42 MHz), this corresponds approximately to 4.8 cm. As shown later, code-phase errors can exceed 100 m. Carrier-phase tracking, thus, yields relatively multipath-free measurements.

b. Region II: Absolute value of the DLL tracking error greater than or equal to one chip. In this region, $R(\tau)$ is approximately zero. Although this is not strictly true, it is a reasonable approximation in light of its proximity to zero relative to the peak value of the correlation function. Using this assumption, Eq. (25) simplifies to the following:

$$\alpha R(\tau-\delta)\sin(\theta_c-\theta_m)=0\tag{27}$$

therefore

$$\theta_c=\theta_m+2\pi N\tag{28}$$

For region II operation, then, the composite phase of the signal entering the PLL is simply the multipath phase relative to the direct signal. The direct signal phase is taken to be zero. Conceptually, region II operation involves the DLL tracking the multipath rather than the direct signal. In either operating region, then, the multipath-induced carrier-phase measurement error is given by θ_c.

C. Noncoherent Pseudorandom Noise Receiver

This section details expressions for the noncoherent DLL discriminator curve in the absence and presence of multipath. Because it can be shown that the PLL carrier-phase error attributable to multipath has the same form as for the coherent receiver architecture, the derivation is not given here.

1. Noncoherent Delay Lock Loop Discriminator Curve in the Absence of Multipath

The signal flow for the noncoherent DLL is similar to that of the coherent DLL except the outputs of the early and late correlators (mixers plus bandpass filters) are squared prior to being low-pass filtered and differenced. The signals entering the LPFs are given by squaring the expressions in Eqs. (7) and (8):

$$\{\overline{s_1(t)s_E(t)}\}^2 = 1/4\, A^2R^2(\tau - \tau_d)\cos^2(\omega_1 t - \theta) \tag{29}$$

$$\{\overline{s_1(t)s_L(t)}\}^2 = 1/4\, A^2R^2(\tau + \tau_d)\cos^2(\omega_1 t - \theta) \tag{30}$$

Note that because the early and late signals are being generated noncoherently, θ is simply the phase offset of the frequency tracking loop. After Eqs. (29) and (30) have been expanded, the LPFs reject the double-frequency terms leaving the following:

$$\{\overline{s_1(t)s_E(t)}\}^2 = 1/8\, A^2R^2(\tau - \tau_d) \tag{31}$$

$$\{\overline{s_1(t)s_L(t)}\}^2 = 1/8\, A^2R^2(\tau + \tau_d) \tag{32}$$

Again, the discriminator function is formed by differencing the outputs of the LPFs. Thus, the normalized form of the noncoherent discriminator function is as follows:

$$D_n(\tau) = R^2(\tau + \tau_d) - R^2(\tau - \tau_d) \tag{33}$$

where the subscript n denotes noncoherent DLL operation.

2. *Noncoherent Delay Lock Loop Discriminator Curve in the Presence of Multipath*

Following a procedure similar to that for the analysis of the coherent DLL, the normalized form of the noncoherent discriminator function is given by the following:

$$\begin{aligned} D_{nm}(\tau) = {} & R^2(\tau + \tau_d) - R^2(\tau - \tau_d) + \alpha^2[R^2(\tau + \tau_d - \delta) - R^2(\tau - \tau_d - \delta)] \\ & + 2\alpha \cos(\theta_m)[R(\tau + \tau_d)R(\tau + \tau_d - \delta) - R,(\tau - \tau_d)R(\tau - \tau_d - \delta)] \end{aligned} \tag{34}$$

where the subscript *nm* denotes noncoherent DLL operation in the presence of a single multipath ray.

D. Simulation Results

Having derived the multipath error equations, simulations allow for the quantification of the error encountered under various multipath conditions.

1. *C/A Code*

The pseudorange error envelope as a function of relative multipath amplitude and delay is given in Fig. 1 for the standard coherent DLL. Note that the standard separation of 1 PRN bit period between the early and late correlators is assumed. In the figure, the relative multipath amplitude is constant and the upper curve represents error attributable to a multipath ray that is in phase with the direct signal. The bottom curve represents the out-of-phase case. It is important to remember that the relative phase of the multipath signal is a function of the electromagnetic properties of the reflecting object in addition to the path length difference. A multipath signal with a given delay can, therefore, take an arbitrary relative phase depending upon the reflector. Examination of the plot reveals that even for a weak multipath signal (−20 dB), the peak error is 15 m. It is also

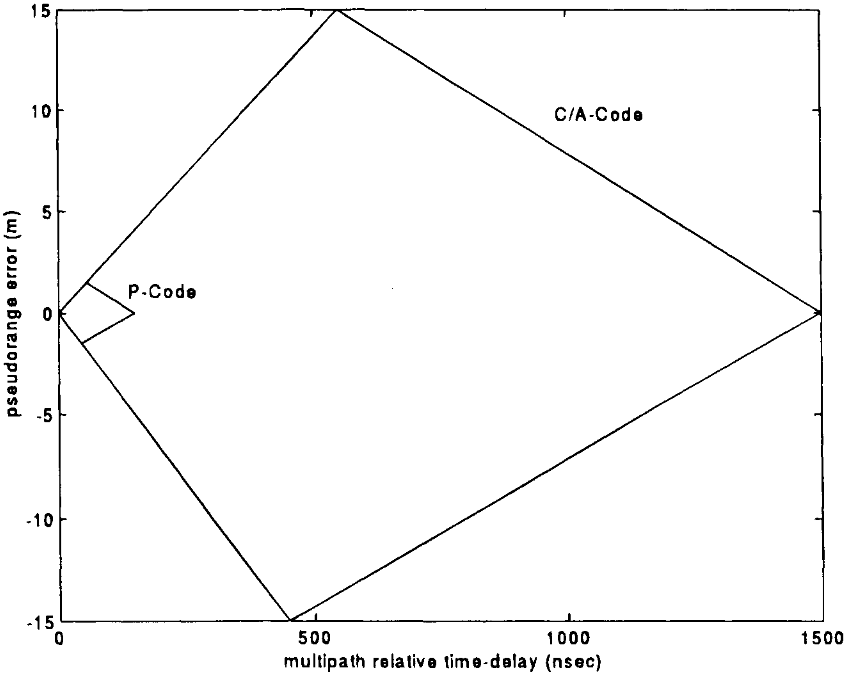


Fig. 1 C/A-code and P-code multipath error envelopes for relative multipath amplitude set at −20 dB.

important to note that the error drops to zero at a delay of 1466 ns (1.5 chips). Furthermore, as shown in Refs. 19 and 20, the error increases when the multipath relative strength increases. Assuming an environment in which the multipath signal strength never exceeds that of the direct, the peak multipath error equals one-half of a chip length.[3] Although not shown, the error envelope for the noncoherent DLL is exactly the same as for the coherent DLL. Their behavior within the envelope is different, however, as is shown later in the plots of average error.

It is important to note that, in reality, multipath signals that are delayed by more than 1.5 chips can induce some error. Recall that in the derivation of the multipath error equations, the sidelobes of the correlation function were ignored. This was not a problem for analyzing multipath with relative time delays of less than 1.5 chips. However, for longer delay multipath, the correlation sidelobes provide the mechanism for interference with the direct signal. The small magnitude of these sidelobes relative to the main lobe means that long delay multipath signals will be attenuated about 24 dB in the correlation process. Although weak, this kind of multipath can induce error on the order of several meters.

For a better understanding of the behavior of the error curve within the envelope, Figs. 2 and 3 show C/A-code pseudorange error over a small range of time delay for different relative multipath amplitudes. For these examples, the multipath relative phase was assumed to be strictly a function of relative path delay. When

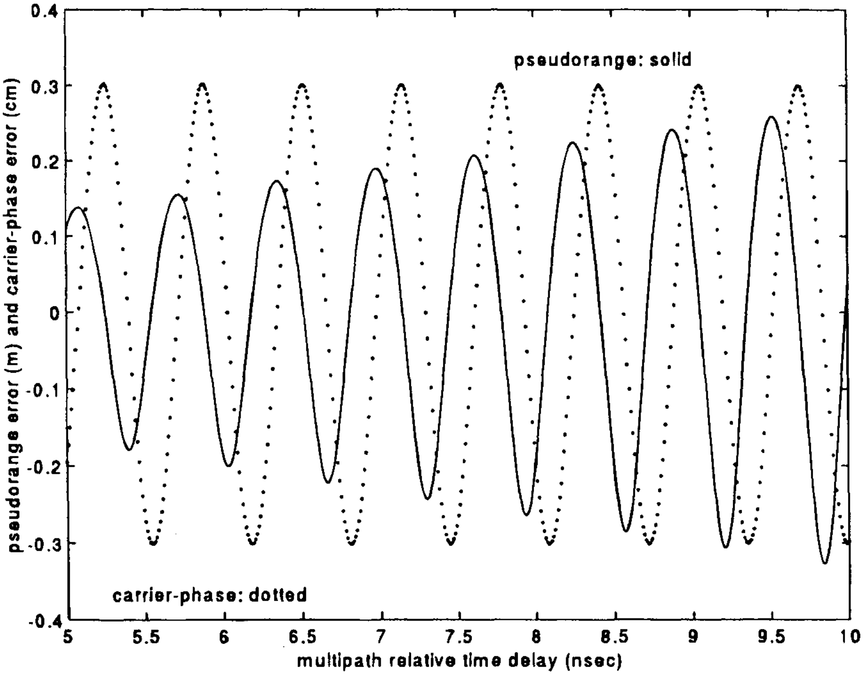


Fig. 2 C/A-code and carrier-phase multipath error for relative multipath amplitude set at −20 dB. Relative phase was calculated strictly from relative path delay.

the relative multipath amplitude is small, the error varies sinusoidally as a function of relative path delay and thus is narrow band. When the relative amplitude is large, however, the error is not sinusoidal and, in fact, it contains sharp discontinuities. In this case, the error signal is wide band.[20,21] This disproves the popular myth that every peak in a multipath error spectrum corresponds to a separate multipath ray. The nonsinusoidal behavior is a result of nonlinearities in the receiver architecture. The effects reach further than simply causing an increased error signal bandwidth. As was first shown by Hagerman[3] and later by Van Nee[17] and Braasch and Van Graas,[21] the error signal is not zero-mean. In situations where the multipath relative phase is fluctuating (nonzero relative phase rate-of-change), the errors will not average out to zero.

Plots of average error vs time delay are given in Fig. 4. Because relative-phase is a complex function of the electromagnetic properties of the reflecting object in addition to the time delay, errors may be computed by holding relative amplitude and phase constant, and varying the time delay. For each of the plots shown, error was averaged over 10 relative-phase values evenly spaced between 0 and 180 deg. As can be seen in the plots, not only are the error traces nonzero mean, the average value can easily be several meters. For example, a relative multipath amplitude of −8 dB yields a peak average error of 8 m for the coherent DLL.

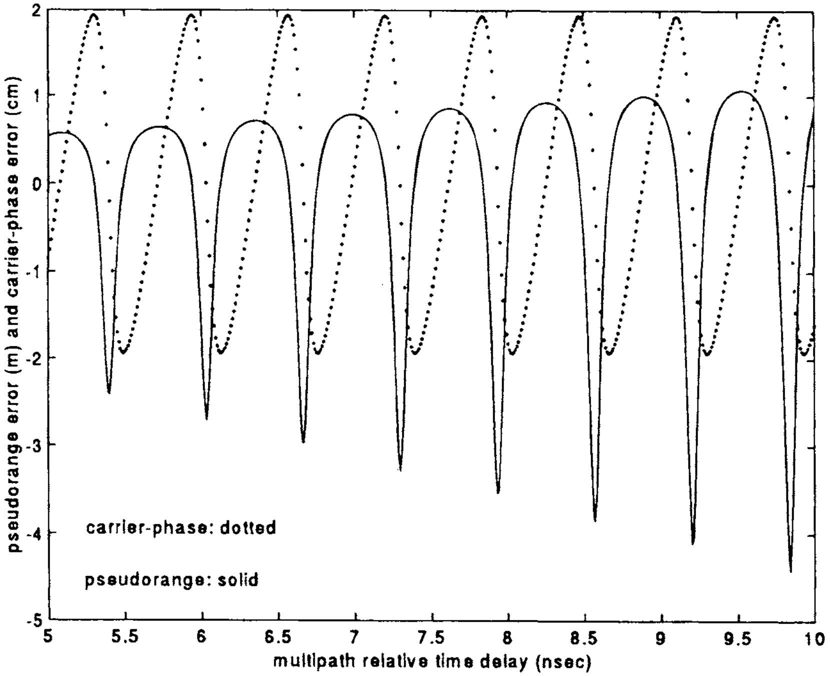


Fig. 3 C/A-code and carrier-phase multipath error for relative multipath amplitude set at −4.4 dB. Relative phase was calculated strictly from relative path delay.

2. *P Code*

The P code is the second PRN code modulated onto the GPS carrier. Although it was made available to the public during the satellite constellation buildup, DOD policy dictates that this code be encrypted (forming the so-called Y code) so as to be accessible only by authorized users. For those users having access to it, however, considerable multipath reduction or rejection is gained. Because the P code (10.23 MHz) is modulated at a rate ten times higher than the C/A code (1.023 MHz), its chips are thus one-tenth the length. The P code, therefore, is much less sensitive because it is affected only by multipath with relative time delays less than 1.5 times its own chip. Instantaneous and average multipath error curves for the P code have exactly the same shape as for the C/A code. In fact, the only difference is the scale factor of 10. The results for the P code may be obtained from Figs. 2–4 simply by dividing the C/A-code pseudorange numbers on both axes by 10. The P-code error envelope is given in Fig. 1.

3. *Carrier Phase*

As mentioned earlier, carrier-phase multipath errors typically are on the order of centimeters. This is illustrated in Figs. 2 and 3. Carrier-phase errors are shown with the corresponding pseudorange errors. It is interesting to note that, as the

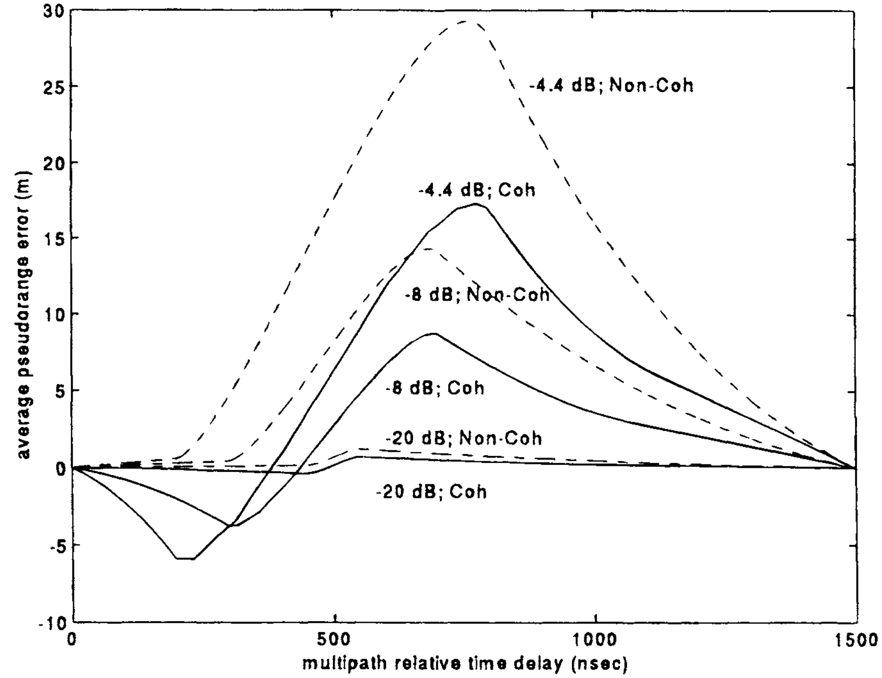


Fig. 4 C/A-code average multipath error. Curves are given for relative multipath amplitudes of −4.4 dB, −8 dB, and −20 dB for both coherent and noncoherent DLLs. For a given time delay for each of the curves, error was averaged over ten relative-phase values evenly spaced between 0 and 180 deg.

theory would predict, the pseudorange errors peak when the carrier-phase errors are zero and vice versa.

III. Aggravation and Mitigation

In the previous section, the basic relationships governing the response of the receiver to multipath were derived. Multipath was parameterized in terms of amplitude, time delay, phase, and phase-rate relative to the direct signal. As was shown, multipath error is proportional to the relative strength of the multipath signal and nonlinearly dependent upon time-delay and phase. The time-delay of a given multipath signal is entirely dependent upon the geometry of the environment in which the receiver is located. The amplitude and phase, however, are dependent upon both the environment and the characteristics of the user's antenna.

A. Antenna Considerations

In Sec. II, the relationships governing the response of the receiver to multipath were derived. Multipath was parameterized in terms of amplitude, time delay, phase and phase-rate relative to the direct signal. Prior to arriving at the receiver tracking loops, multipath must pass through the antenna. Realizable antennas (as opposed to theoretical isotropic radiators) do not receive signals equally from all directions. In fact, partial multipath rejection is built in to some antenna designs by shaping the gain pattern. Because most multipath arrives from angles near the horizon, multipath may be reduced by shaping the pattern to have low gain in these directions. However, extensive shaping of the pattern requires either a large aperture or multiple elements and signal processing. This might be feasible for the reference station in a differential system. However, this is not possible for highly dynamic platforms (i.e., aircraft), which require compact antennas and virtually omnidirectional gain.

Additional multipath attenuation by the antenna results from polarization discrimination. The direct GPS signal incident on the antenna is right-hand circularly polarized. In general, a single reflection from a planar surface will be left-hand elliptically polarized if the angle of incidence is less than the Brewster angle.[22] An ideal GPS antenna would completely reject all signals that are left-hand circularly polarized. In reality, total rejection is not obtained, but attenuation on the order of 10 dB is typical. The situation deteriorates further when we consider reflectors with nonsmooth surfaces. Reflections from very rough surfaces have random polarization characteristics. As a result, typically only 3 dB of attenuation can be achieved.

Having determined the multipath rejection properties of the antenna itself, the issue of antenna siting can be addressed. The multipath error equations (discussed in Sec. II) along with the antenna characteristics (just described) give general guidance regarding the siting of ground reference stations. To minimize the effects of multipath, the antenna should be located as far away from other objects as possible and/or should be located in such a way that multipath arrives from directions in which the antenna has low gain.

Locating the antenna away from objects minimizes signal blockage and reduces the strength of multipath signals. Mere distance, however, does not guarantee multipath error immunity. One of the great GPS multipath myths is the following (C/A-code processing assumed): GPS receivers reject multipath signals that arrive more than 1.5 chips (440 m or 1466 ns) after the direct signal; therefore, objects greater than 440 m from the receiver are not of concern. The error in this statement stems from a misinterpretation of the concept of multipath delay. As shown in sec. II, the 1.5 chips delay is a delay of the multipath *relative to the direct signal.* As was shown in Braasch,[20] the relative delay of the multipath is a function of receiver-to-object distance *and* proximity of the object to the satellite-to-receiver line-of-sight. The closer an object is to the line-of-sight, the greater the receiver-to-object distance can be while still yielding a relative multipath delay of less than 1.5 chips. The conclusion, then, is that increased receiver-to-object distance only guarantees a weakening of the multipath signal.

Maximization of the receiver-to-object distance was the first general siting guideline mentioned above. The second considered the gain pattern of the antenna. Multipath mitigation can be maximized if the antenna is located so that multipath arrives from directions in which the antenna has low gain. Multipath usually arrives from angles near the horizon and below. Multipath mitigation, then, is achieved by using an antenna with low gain in these directions.

B. Receiver Design

As demonstrated in Sec. II, multipath error is highly dependent upon receiver architecture. The following sections expand upon this.

1. Coherent vs Noncoherent Delay Lock Loop Architectures

As the plots in Sec. II demonstrate, for a given set of multipath parameters, average multipath error is smaller for a coherent DLL than for a noncoherent DLL. Furthermore, van Nee[19] has shown that when the relative phase rate-of-change of the multipath is large compared to the code tracking loop bandwidth, the coherent DLL error does average to zero. The primary disadvantage of the coherent DLL is a matter of robustness and not multipath. A coherent DLL will be disrupted by cycle slips, whereas a noncoherent DLL will continue to function independently of the carrier-phase tracking loop.

2. C/A Code with Narrow Correlator Spacing

Fenton et al.[23] and Van Dierendonck et al.[24] describe a patented development in GPS receiver design that lessens the effect of multipath by narrowing the spacing of the early and late correlators in a noncoherent DLL. By using a small portion of the correlation function (around the peak) to form the discriminator, maximum multipath error is reduced by a factor of 10 and multipath with relative delays of approximately 1 chip or greater is rejected entirely (0.1 chip correlator spacing). Braasch[26] and Van Nee[18] showed that the narrow correlator concept applies to the coherent DLL also. Meehan and Young[25] describe variations of the narrow correlator concept which benefit carrier-phase measurements as well as code measurements.

3. P Code

As discussed in Sec. II the P-code multipath performance is far superior to C/A-code tracking. For a given set of multipath parameters, peak error on the C/A code is 10 times larger than on the P code. When access is available, P code is clearly the choice when multipath is considered. It is important to note, however, that the 10:1 multipath error reduction applies to the maximum error and not necessarily the instantaneous error. This is illustrated in Fig. 1 where a relative multipath amplitude of -20 dB has been simulated. Note that for short delay multipath, the P-code and C/A-code multipath error envelopes coincide. Thus, for multipath in this region, the P code will not necessarily perform an order of magnitude better than the C/A code.

IV. Multipath Data Collection

In many instances it is desirable to evaluate the multipath in a given environment. This is particularly true in the siting of a ground reference station. Thus, the situation arises that GPS data are to be collected, and the multipath component of the error is to be isolated. Note that a truth reference system is not useful in this case. A truth reference system allows for the determination of the total system error. However, in this case only the multipath component of the error is desired.

The GPS signal itself may be exploited to isolate the combination of code multipath plus receiver error[4,11]. The GPS code and carrier-phase (integrated Doppler) measurements may be expressed as follows:[26]

$$\rho_{\text{code}} = D + c(b_u - B) + c(T + I + M_{\text{code}} + \text{HW} + \nu_{\text{code}}) + \text{URE} + \text{SA} + \text{MEAS}_{\text{code}} \tag{35}$$

$$\rho_{\text{phase}} = D + c(b_u - B) + c(T - I + M_{\text{phase}} + \text{HW} + \nu_{\text{phase}}) + \text{URE} + \text{SA} + \text{MEAS}_{\text{phase}} + \Delta \tag{36}$$

where

B = satellite clock offset from system time

b_u = receiver clock offset from system time

c = speed of light, m/s

D = true line-of-sight range from the satellite to the user in meters

HW = receiver hardware delay, s

I = apparent change in signal path delay due to propagation through the ionosphere, s

M = apparent change in signal path delay due to specular (i.e., nondiffuse) multipath, s

MEAS = receiver measurement (tracking) errors

SA = (selective availability) range error due to the intentional degradation of the satellite clock and orbit information by the DOD (used for security purposes), m

T = apparent change in signal path delay due to propagation through the troposphere, s

URE = (user range error) range error due to satellite clock and orbit uncertainty, m

ν = apparent change in signal path delay due to a combination of receiver noise and diffuse multipath, s

Δ = range difference between the code and integrated Doppler measurements due to an integer wavelength ambiguity, m

ρ_{code} = code measurement (pseudorange), m

ρ_{phase} = carrier-phase (integrated Doppler) measurement, m

Diffuse multipath arises from reflection and diffraction from a group of electrically small objects. Each of these objects individually produces a negligible multipath field, but the sum effect can be on the order of receiver noise values. This effect is lumped in with receiver noise because it is generally uncorrelated over time, and therefore noise-like in behavior. Note that the ionosphere term is equal in magnitude but opposite in sign for the two measurements. More detailed information on the effects of the ionosphere can be found in Chapter 12, this volume. The integer wavelength ambiguity arises in the carrier measurement because the basic measurement is that of integrated Doppler (change-of-range relative to the start of signal tracking). For the moment, consider the observable obtained by differencing the code and carrier-phase measurements:

$$\rho_{\text{code}} - \rho_{\text{phase}} = 2I + \text{MEAS}_{\text{code}} - \text{MEAS}_{\text{phase}} + \nu_{\text{code}} - \nu_{\text{phase}} + M_{\text{code}} - M_{\text{phase}} - \Delta \tag{37}$$

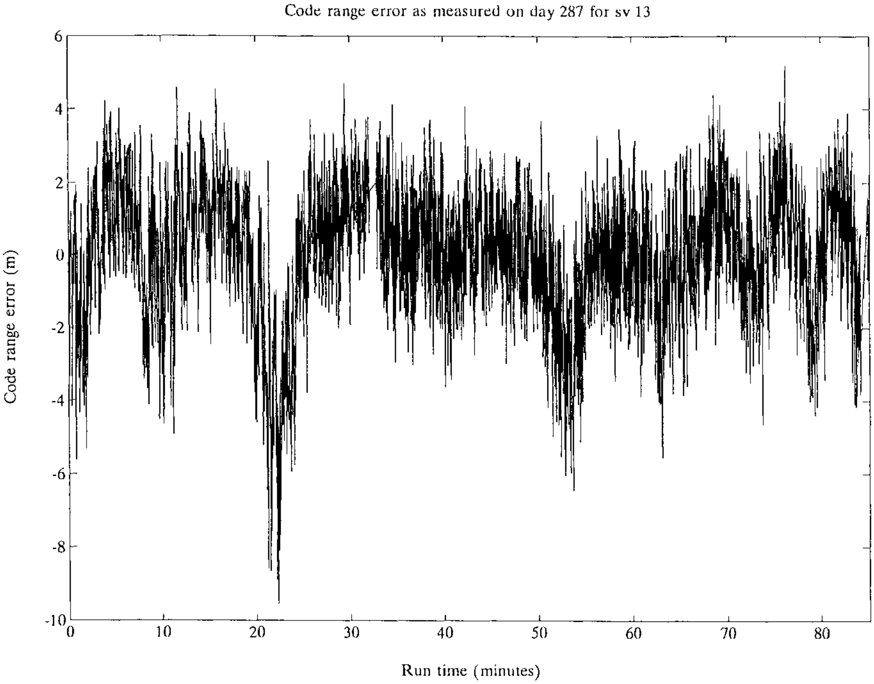


Fig. 5 Ohio University hangar test. Data were collected using an Ashtech 3DF receiver approximately 4 m from the hangar (which is approximately 10 m tall). Note that the curve shows the combined effects of multipath and receiver noise.

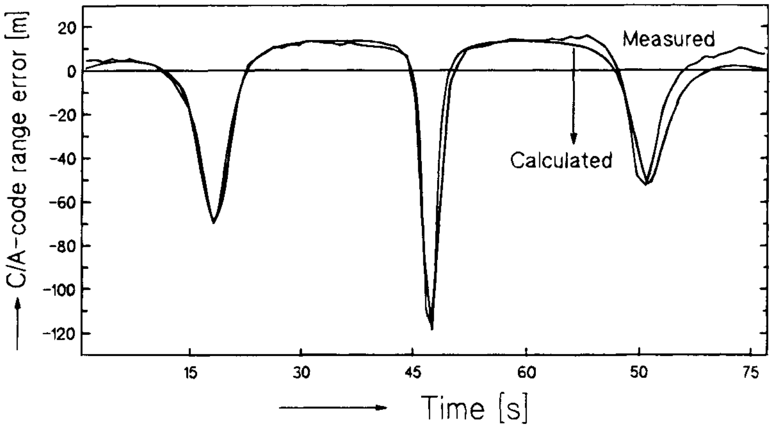


Fig. 6 Delft University electrical engineering building test. Data were collected using a Trimble 4000SST receiver approximately 30 m from the building (which is approximately 100 m tall). The curve labeled "calculated" is the result of fitting a model based on theory described in Sec. II. Reprinted from Van Nee (1992) with permission from R. van Nee and the Delft University of Technology.

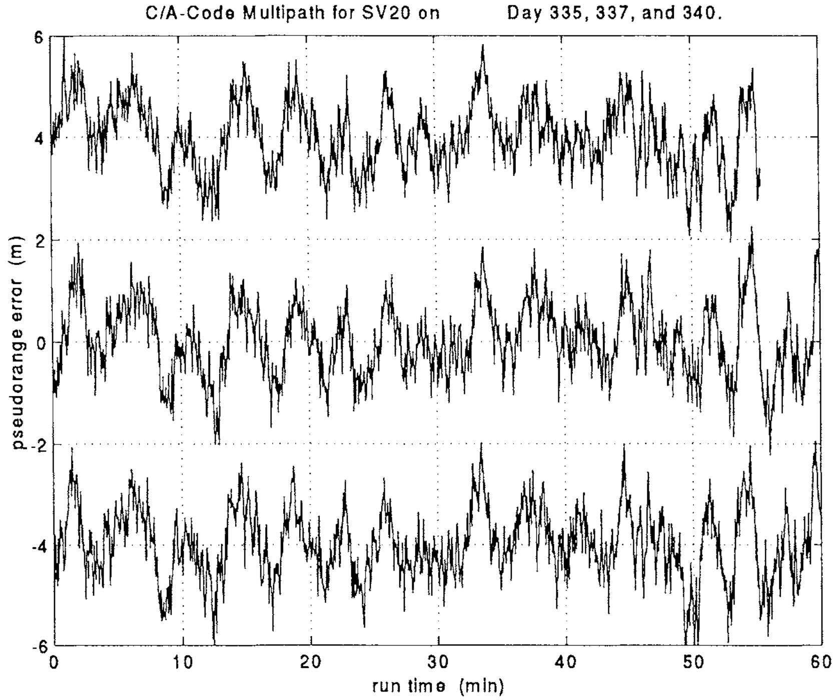


Fig. 7 Day-to-day repeatability of C/A-code multipath. Curves show pseudorange multipath error for three days. The data from day 235 have been offset in amplitude by +4 m and in time by 0 min; data from day 237 have been offset in amplitude by 0 m and in time by 8 min; and data from Day 240 have been offset in amplitude by −4 m and in time by 20 min. Data were collected using a Turborogue receiver on the roof of the University of Colorado, Boulder, Engineering Center.

Phase measurement errors, phase noise, and multipath typically are negligible.[27,28] The usual application of this process involves differencing code and carrier data collected over a given period of time. The integer ambiguity may, thus, be removed by subtracting out the bias. What remains is a combination of multipath, receiver code measurement error, noise, and an ionospheric term:

$$(\rho_{\text{code}} - \rho_{\text{phase}})_{\text{adj}} = 2I + \text{MEAS}_{\text{code}} + \nu_{\text{code}} + M_{\text{code}} \tag{38}$$

Receiver code measurement errors typically result from dynamics-induced tracking loop lags.[29] Although these are usually correlated over time, receiver noise is not, and therefore it may be reduced through filtering. The filtering scheme smooths the code against the carrier measurements. Typical implementations are the Hatch filter[30] and the complementary Kalman filter.[31] The remaining term besides multipath is then the effect due to the ionosphere. However, measurements from two different carrier frequencies can be used to eliminate the ionospheric term. For those users with single-frequency receivers, the effect of the ionosphere is typically a long-term drift, which can easily be distinguished from the higher-frequency multipath errors.

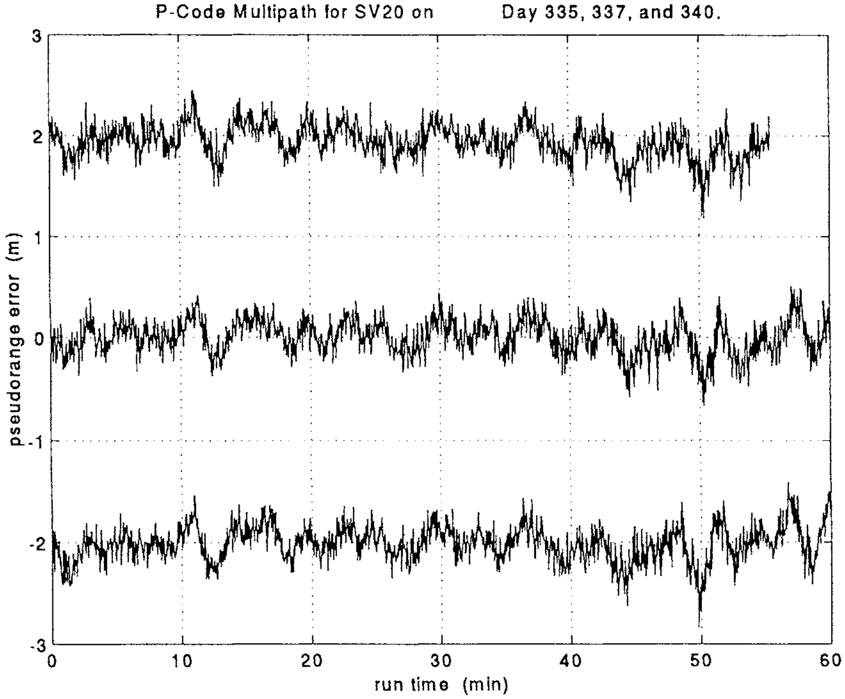


Fig. 8 Day-to-day repeatability of P-code multipath. Curves show pseudorange multipath error for three days. The data from Day 235 have been offset in amplitude by +2 m and in time by 0 min; data from Day 237 have been offset in amplitude by 0 m and in time by 8 min; and data from Day 240 have been offset in amplitude by −2 m and in time by 20 min. Data were collected using a Turborogue receiver on the roof of the University of Colorado, Boulder, Engineering Center.

As a result, the differencing of code and carrier measurements allows for the isolation of pseudorange multipath and receiver code measurement errors. The magnitude and behavior of receiver code measurement errors is, quite obviously, receiver dependent. However, it must be taken into consideration when analyzing the data. Test results indicate that receiver code measurement errors can be on the order of one meter.

For static data collection efforts, the repeating pattern of the satellite orbits can be exploited to gain confidence in the multipath analysis. The question often arises as to whether a given set of error data is truly multipath or not. The answer lies in collecting data on successive days. If, for instance, data are collected from satellite 20 on Wednesday between 2:00 p.m. and 3:00 p.m., data from the same satellite should be collected during the same time on Thursday. Because the satellite orbits the Earth twice every sidereal day (23 h 56 min), it returns to the same location four minutes earlier each day. Thus, errors thought to be multipath can be checked for repeatability. Any residual noise errors and the like will not repeat from day to day. However, because the same antenna, object, and satellite positions are being used, the multipath data should repeat. In this way, confidence

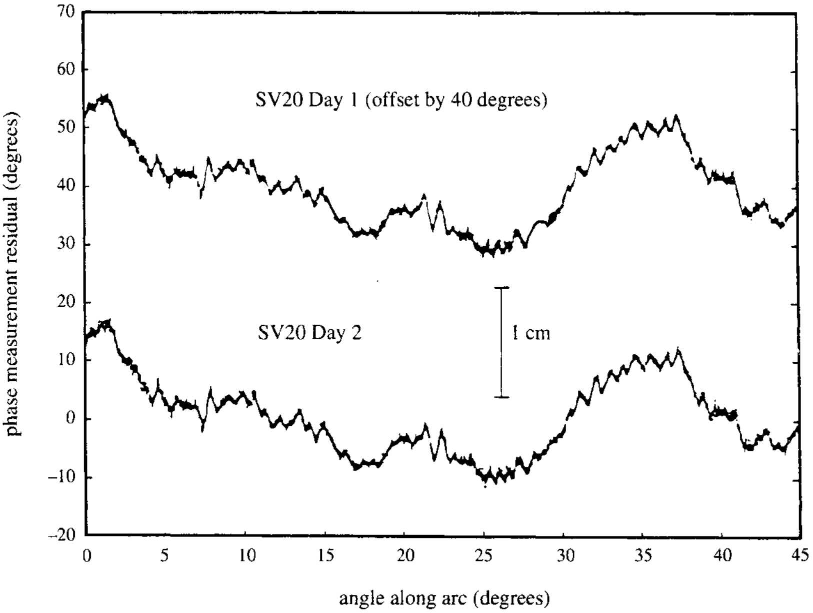


Fig. 9 Carrier-phase multipath error. Data was collected using a Trimble Navigation TANS-based GPS receiver. Multipath was induced by mounting a rectangular sheet metal reflector near one of the interferometer antennas. Reprinted from Cohen (1992) with permission.

can be gained that a set of error data is truly due to multipath. Obviously, this method assumes that the multipath comes from stationary objects.

Examples of GPS code multipath errors are given in Figs. 5–8. For the data shown in Fig. 5, a GPS antenna was placed approximately 4 m in front of a hangar door at the Ohio University airport.[32] In this case, the measurements were not filtered, and thus the graph shows a combination of multipath and receiver noise on the code measurements. The low-frequency trend shows the oscillating error attributable to the phase variation of the multipath. The trend is much clearer in fig. 6. These data were collected by the Electrical Engineering Department at the Delft University of Technology in Delft, The Netherlands.[18] A GPS antenna was placed approximately 30 m away from the Delft University Electrical Engineering building. The building is approximately 100 m high and 70 m wide. This figure illustrates several points about multipath error. First, it is highly nonsinusoidal. This verifies the earlier claim and is a direct result of the nonlinearities in the receiver architecture. Second, the maximum error is −120 m, thus verifying the assertion that code multipath errors can be extremely large. Finally, using the equations described in Sec. II, a model was fit to the data with the assumption that there were two major reflectors. Time delays were calculated from the geometry of the test set-up. The other parameters were adjusted to achieve a good fit. The results show clearly that the theoretical equations are good descriptions of actual multipath error.

Figures 7 and 8 illustrate C/A-code and P-code multipath day-to-day repeatability. The data sets were collected on DOY 335, 337, and 340 of 1993 (Dec. 1, 3, 6) on the roof of the Engineering Center at the University of Colorado, Boulder. Each of the three days of data for each code has been offset from each other for visual clarity (4 m for the C/A code, 2 m for the P code). The lower frequency oscillations due to multipath are apparent as well as the day-to-day repeatability. Sample correlation coefficients between data sets are on the order of 0.50–0.60.

Figure 9 illustrates carrier-phase multipath and its repeatability in a controlled environment. These data were collected at Stanford University using a short baseline GPS interferometer.[33] A metal sheet was intentionally placed near one antenna to induce multipath. The carrier-phase residuals clearly demonstrate the presence of multipath.

Acknowledgments

The author gratefully acknowledges the assistance of the Delft University of Technology, Faculty of Electrical Engineering, Telecommunications and Traffic-Control Systems Group (Delft, The Netherlands). In particular, the help of Edward Breeuwer, Richard van Nee, and Durk van Willigen was invaluable. Penina Axelrad and Christopher Comp of the Colorado Center for Astrodynamics Research, University of Colorado, Boulder, are thanked for their comments on the manuscript and for the collection of the C/A-code and P-code multipath data. Clark Cohen, Department of Aeronautics and Astronautics, Stanford University is thanked for the provision of the carrier-phase multipath error plot.

References

[1]Lachapelle, G., Falkenberg, W., Neufeldt, D., and Keilland, P., "Marine DGPS Using Code and Carrier in a Multipath Environment," *Proceedings of ION GPS-89,* Colorado Springs, CO, Sept. 1989, Institute of Navigation, Washington, DC, pp. 343–347.

[2]Cohen, C., and Parkinson, B., "Mitigating Multipath Error in GPS-Based Attitude Determination," *Guidance and Control 1991,* Vol. 74, *Advances in the Astronautical Sciences,* edited by R. Culp and J. McQuerry, American Astronautical Society, 1991, pp. 53–68.

[3]Hagerman, L., "Effects of Multipath on Coherent and Noncoherent PRN Ranging Receiver," Aerospace Rep. TOR-0073(3020-03)-3, Development Planning Division, The Aerospace Corporation, May 15, 1973.

[4]General Dynamics, Electronics Division, "Final User Field Test Report for the NAVSTAR Global Positioning System Phase I, Major Field Test Objective No. 17: Environmental Effects, Multipath Rejection," Rept. GPS-GD-025-C-US-7008, San Diego, CA, March 28.

[5]Counselman, C., and Gourevitch, S., "Miniature Interferometer Terminals for Earth Surveying: Ambiguity and Multipath with Global Positioning System," *IEEE Transactions on Geoscience and Remote Sensing,* Vol. GE-19, No. 4, 1981, pp. 244–252.

[6]Counselman, C., "Miniature Interferometer Terminals for Earth Surveying (MITES): Geodetic Results and Multipath Effects," Digest of the International Geoscience and Remote Sensing Symposium, Washington, DC, June 1981, Institute of Electrical and Electronics Engineers, New York, pp. 219–224.

[7]Greenspan, R., Ng, A., Przyjemski, J., and Veale, J., "Accuracy of Relative Positioning by Interferometry with Reconstructed Carrier GPS: Experimental Results," *Proceedings of the Third International Geodetic Symposium on Satellite Doppler Positioning,* New Mexico State University, Las Cruces, NM, Feb. 1982, pp. 1177–1195.

[8]Bletzacker, F., "Reduction of Multipath Contamination in a Geodetic GPS Receiver," Proceedings of the First International Symposium on Precise Positioning with the Global Positioning System, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, Rockville, MD, April 1985, pp. 413–422.

[9]Tranquilla, J., and Carr, J., "GPS Multipath Field Observations at Land and Water Sites," *Navigation,* Vol. 37, No. 4, 1990–1991, pp. 393–414.

[10]Falkenberg, W., Kielland, P., and Lachapelle, G., "GPS Differential Positioning Technologies for Hydrographic Surveying," Record of the Position, Location, and Navigation Symposium, PLANS, Orlando, FL, Institute of Electrical and Electronics Engineers, New York, Dec. 1988, pp. 310–317.

[11]Evans, A., "Comparison of GPS Pseudorange and Biased Doppler Range Measurements to Demonstrate Signal Multipath Effects," Proceedings of the International Telemetering Conference, Las Vegas, NV, Instrument Society of America, Research Triangle Park, NC, Oct. 1986, pp. 795–801.

[12]Georgiadou, Y., and Kleusberg, A., "On Carrier Signal Multipath Effects in Relative GPS Positioning," Manuscripta Geodaetica, Vol. 13, 1988, pp. 172–179.

[13]Abidin, H., "Extrawidelaning for 'On the Fly' Ambiguity Resolution: Simulation of Multipath Effects," *Proceedings of ION GPS-90,* Colorado Springs, CO, Institute of Navigation, Washington, DC, Sept. 19–20, 1990, pp. 525–533.

[14]Bishop, G., Klobuchar, J., and Doherty, P., "Multipath Effects on the Determination of Absolute Ionospheric Time Delay From GPS Signals," *Radio Science,* Vol. 20, No. 3, 1985, pp. 388–396.

[15]Sennott, J., and Pietraszewski, D., "Experimental Measurement and Characterization of Ionospheric and Multipath Errors in Differential GPS," *Navigation,* Vol. 34, No. 2, 1987, pp. 160–173.

[16]Sennott, J., and Spalding J., "Multipath Sensitivity and Carrier Slip Tolerance of an Integrated Doppler DGPS Navigation Algorithm," Record of the Position Location and Navigation Symposium (PLANS), Las Vegas, NV, Institute of Electrical and Electronics Engineers, New York, March 1990, pp. 638–644.

[17]van Nee, R., "Multipath Effects on GPS Code Phase Measurements," *Proceedings of ION GPS-91,* Albuquerque, NM, Institute of Navigation, Washington, DC, Sept. 1991, pp. 915–924.

[18]van Nee, R., "GPS Multipath and Satellite Interference," *Proceedings of the Forty-eighth Annual Meeting of the Institute of Navigation,* Institute of Navigation, June 1992, Washington DC, pp. 167–177.

[19]van Nee, R.,"Spread Spectrum Code and Carrier Synchronization Errors Caused by Multipath and Interference," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 29, No. 4, 1993, pp. 1359–1365.

[20]Braasch, M., "On the Characterization of Multipath in Satellite-Based Precision Approach and Landing Systems," Ph.D. Dissertation, Department of Electrical and Computer Engineering, Ohio University, Athens, OH, June 1992.

[21]Braasch, M., and van Graas, F., "Mitigation of Multipath in DGPS Ground Reference Stations," *Proceedings of the ION National Technical Meeting,* San Diego, CA, Institute of Navigation, Washington, DC, Jan. 1992, pp. 105–114.

[22]Balanis, C., *Advanced Engineering Electromagnetics,* Wiley, New York, 1989, pp. 185–196.

[23]Fenton, P. et al., "Novatel's GPS Receiver: The High Performance OEM Sensor of the Future," *Proceedings of ION GPS-91,* Albuquerque, NM, Institute of Navigation, Washington, DC, Sept. 1991, pp. 49–58.

[24]Van Dierendonck, A., Fenton, P., and Ford, T., "Theory and Performance of Narrow Correlator Spacing in a GPS Receiver," *Navigation,* Vol. 39, No. 3, 1992, pp. 265–283.

[25]Meehan, T., and Young, L., "On-Receiver Signal Processing for GPS Multipath Reduction," Proceedings of the 6th International Geodetic Symposium on Satellite Positioning, Columbus, OH, Defense Mapping Agency and the Ohio State University, Columbus, OH, March 1992, pp. 200–208.

[26]Braasch, M., "A Signal Model for GPS," *Navigation,* Vol. 37, No. 4, 1990–1991, pp. 363–377.

[27]Ferguson, K., et al., "Three-Dimensional Attitude Determination with the Ashtech 3DF 24-Channel GPS Measurement System," *Proceedings of the ION National Technical Meeting,* Phoenix, AR, Institute of Navigation, Washington, DC, Jan. 1991.

[28]Braasch, M., and van Graas, F., "Guidance Accuracy Considerations for Real-Time GPS Interferometry," *Proceedings of ION GPS-91,* Albuquerque, NM, Institute of Navigation, Washington, DC, Sept. 1991, pp. 373–386.

[29]Braasch, M., "Isolation of GPS Multipath and Receiver Tracking Errors," *Proceedings of the ION National Technical Meeting,* San Diego, CA, Institute of Navigation, Washington, DC, Jan. 1994, pp. 511–521.

[30]Hatch, R., "The Synergism of GPS Code and Carrier Measurements," *Proceedings* of the Third International Geodetic Symposium on Satellite Doppler Positioning, New Mexico State University, Las Cruces, NM, Feb. 1982, pp. 1213–1231.

[31]van Graas, F., and Braasch, M., "GPS Interferometric Attitude and Heading Determination: Initial Flight Test Results," *Navigation,* Vol. 38, No. 4, 1991–1992, pp. 297–316.

[32]Breeuwer, E., "Modeling and Measuring GPS Multipath Effects," Master's Thesis, Faculty of Electrical Engineering, Delft University of Technology, Delft, The Netherlands, Jan. 1992.

[33]Cohen, C., "Attitude Determination Using GPS," Ph.D. Dissertation, Department of Aeronautics and Astronautics, Stanford University, Stanford, CA, Dec. 1992.

Chapter 15

Foliage Attenuation for Land Mobile Users

J. J. Spilker Jr.*
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

LAND mobile users are expected to be one of the largest categories of GPS users, and it is important to examine the specific GPS propagation issues for this environment. It has already been shown that the optimum geometric dilution of precision (GDOP) is provided when several of the GPS satellites are at low elevation angle near the horizon. However, the land mobile user environment differs from that of aircraft in flight or ships at sea in that the user driving along a road or freeway is often subject to shadowing, diffraction, and scattering of the satellite signal by trees, utility poles, buildings, or hills. These effects are accentuated by the need for operation at low elevation angles for at least some of the GPS satellites. In addition, the requirement for receiver simplicity, and the need to track several satellites widely spaced in angle simultaneously, generally dictates the use of an omnidirectional or hemispherical antenna. Thus, while receiving the direct line-of-sight ray from the satellite, the user has little means to discriminate against multipath signals scattered from ground reflections, tree limbs and foliage, or other scattering elements. In addition, the direct ray may itself be attenuated by tree foliage.

This chapter describes and models the statistics of the signal environment for the rural land mobile user where tree foliage often shadows the user. The statistical models are then compared with measured data for *L*-band signals at various elevation angles. Both stationary and mobile users are considered.

The satellite-to-land mobile user links can be categorized in one of three forms: 1) line of sight, unshadowed; 2) shadowed by trees or foliage; and 3) completely obstructed where the link is totally blocked by hills or other major obstructions (see Fig. 1).

The unshadowed link is characterized by a link where there is a clear line of sight to the satellite, but there can also be a scattered and/or specular multipath component in addition to the line-of-sight link. The multipath component can be caused by ground reflections. The unshadowed link can be modeled as the sum

* Ph.D., Chairman of the Board.

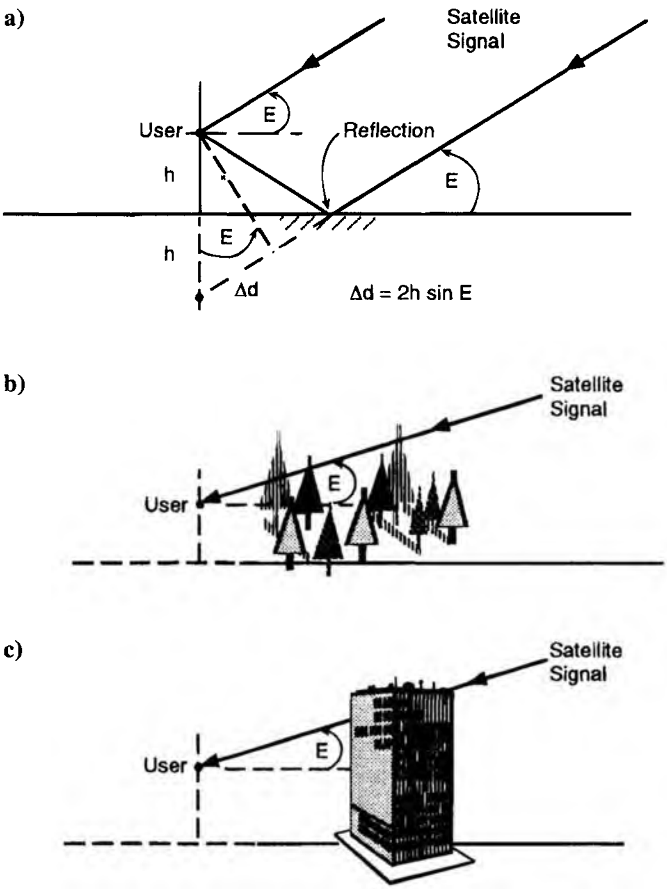


Fig. 1 Different types of links experienced by a mobile user, or fixed user on the ground: a) unshadowed link with multipath reflection a distance *h* below the user and satellites at elevation angle *E*; b) link partially shadowed by trees, telephone poles; c) link completely shadowed by buildings, hills, or other obstructions.

of a fixed line-of-sight signal plus a Rayleigh scattered term. The sum of these two components has a distribution alternatively known as a Rician or a Nakagami–Rice distribution.[1]

Partially shadowed links have a randomly attenuated direct link in addition to the scattered component. The attenuation generally is caused by individual trees, utility poles, clusters of trees, or forests along a roadway. The shadowed signal is sometimes modeled by assuming a lognormal distribution for the direct ray. The sum of direct and scattered components has a Rician distribution for any given amplitude direct ray; i.e., the conditional distribution for a given line-of-sight signal amplitude is Rician. The resulting distribution, obtained by averaging over the lognormal direct ray, is known as the Loo distribution.[2]

The completely obstructed link may be blocked by hills or large man-made structures and has no useful signal component at all over a substantial period of time.

Discussion in this chapter is limited to the shadowed/unshadowed environment common along road/highways in the rural or suburban environment. Typically, a vehicle moves from the shadowed environment where forests are adjacent to the roadway to regions of relatively open space with line-of-sight propagation and then back to the shadowed region. In a given region, there is a certain fraction of time S that a mobile user is in the shadowed region, and during the remaining fraction of time $1-S$ the user is in the unshadowed region.

The channel model employed in this chapter assumes that the received signal $S_r(t)$ has the form

$$S_r(t) = \alpha(t)s(t-\tau) + \Sigma\beta_i(t)s(t-\tau-\Delta\tau_i) \tag{1}$$

where $\alpha(t)$ represents the random path attenuation caused by the foliage, τ is the delay of the direct line-of-sight satellite user path, and $\beta_i(t)$ represents the random attenuation of the scattered components with additional delay $\Delta\tau_i$. For purposes of this chapter, the delays $\Delta\tau_i$ of the scattered components are considered negligible compared to the inverse bandwidth of the signal. Thus, the channel considered here is strictly limited to the flat fading channel; i.e., the signal spectrum is unchanged by the fading except for a random amplitude. Larger multipath delays; for example, as experienced by an aircraft in flight with sea surface reflections, could produce frequency selective fading. Multipath channels where the delays are not negligible are considered in another chapter.

The discussion that follows first treats the stationary user and views the various GPS satellites through individual trees or groves of trees at different elevation angles. The remainder of the chapter emphasizes mobile users where the link is varying fairly rapidly with time as the vehicle moves past trees and other short duration obstructions.

II. Attenuation of an Individual Tree or Forest of Trees—Stationary User

Consider first the attenuation caused by an individual tree where a GPS satellite is viewed by a stationary user at an elevation angle E deg. Goldhirsh and Vogel[1,3] have published measurements of attenuation of an individual tree vs. elevation angle. The approximate geometry of the tree and the relative position of the receiver are shown in Fig. 2. These measurements were taken at 870 MHz, but the results can be translated to L-band after being scaled up to 1.575 GHz using the following relationship:

$$\begin{gathered}\text{Attenuation } (f) \sim \sqrt{f} \\ \text{Attenuation (1.575 GHz) dB} \approx \sqrt{\frac{1575\text{ MHz}}{870\text{ MHz}}}\ \text{Attenuation (870 MHz) dB}\end{gathered} \tag{2}$$

where the attenuation in dB is scaled by $\sqrt{1575/870} = 1.345$. That is, foliage attenuation varies approximately as the square root of frequency.*

*The square root relationship is only a rough approximation. In the frequency range 800 MHz to 2 GHz the rms error has been estimated[1] at 6%.

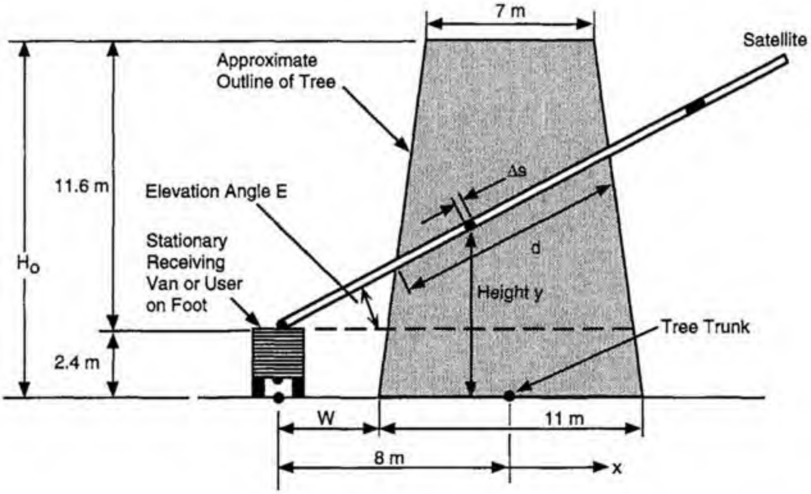


Fig. 2 Simplified geometric configuration of a single callery pear tree vs elevation angle in measurements made by Goldhirsh and Vogel[3] as viewed by a stationary user. The origin of the *x*, *y* coordinates used in the text is the base of the tree trunk.

It has also been shown that the attenuation of a tree in full foliage is roughly 35% greater in dB than that of a deciduous tree without foliage. Thus, the bulk of a tree's attenuation is clearly caused by the wood tree limbs, branches, and trunk rather than by the leaves.

Foliage attenuation is often characterized as attenuation in dB/m of foliage penetration. As shown later, these numbers of attenuation per meter tend to vary widely depending on the type of tree and the height of the ray relative to the top of the tree because the foliage density varies with height. The attenuation also varies with the distance of the tree from the user. Fresnel diffraction analyses have been used to explain the average decrease of attenuation with increasing distance of the receiver to the tree.[4]

As an attempt to obtain a simple purely empirical geometric interpretation of the foliage attenuation refer to Fig. 2 again. Consider a ray path through a nonuniform density foliage slab where for each incremental distance Δs along the ray path has some random value of attenuation in dB/m. The mean attenuation coefficient at that position depends on various parameters, such as type of tree, density of trees, position relative to center, and base of tree. The total attenuation then depends on the distance W from the receiver to the edge of the trees, the elevation angle to the satellite, and the geometric extent of the foliage. In effect, we consider the trees to be a slab of foliage with nonuniform attenuation where the attenuation density in dB/m varies with position. The quantity H_o is the height of the foliage slab representing the tree or trees. It is assumed that $A_d(x, y)$ represents the mean attenuation density vs position, and the actual attenuation is either 0 or k_o depending on whether that position in space has foliage or not. The probability of the incremental ray being attenuated at that position cell is p, and all cells are independent. Thus, the mean attenuation for that incremental cell of length Δs is pk_o where k_o varies with position relative to the base of each tree.

It should be emphasized that this representation is only a simple empirical model which we can attempt to fit to the measured data and compare with measured data. If we use the model for attenuation just discussed and we use the law of large numbers for a large number of independent attenuation cells of length Δs, the total path attenuation in dB approximately follows a normal distribution. The distribution of received signal level in absolute terms then follows a lognormal distribution. We return to this distribution later.

We can easily show that the total distance d through the foliage of the model of Fig. 2 does not change substantially as we vary the elevation angle from 0 to 45 deg; d remains at approximately 10 m. Consequently, a uniform attenuation density foliage slab model would show little variation in total attenuation with elevation angle. However, the attenuation measured by Ref. 1 varies considerably from approximately 9.5 dB at 40 deg to approximately 19 dB at 20 deg elevation angle (measured numbers are scaled up to 1.545 GHz). Thus, the attenuation density in dB/m obviously must vary substantially with position relative to this base of the tree if the attenuation vs elevation angle of the model is to match the measurement. That variation is consistent with what we might expect from a simple observation of the wood portion of a typical tree, which has a greater wood density near the base of the trunk.

We have selected an empirical attenuation density profile (in dB/m) that varies with position relative to the base of the tree's center (the trunk), which seems to match fairly well-measured data for the tree of Fig. 2. This attenuation profile is shown in Fig. 3 for this given tree and tree–user distance. The mean attenuation density in dB/m of this empirical model varies as follows:

$$A_d(x, y) = (1 - y/14)^2(1 - |x|/5.5)\text{K dB/m} \tag{3}$$

This attenuation density profile can be integrated along the path s through the tree at a given elevation angle E to obtain the total attenuation $A_T\,(E)$ where the following obtains:

$$A_T(E) = \int_{\text{path}} A_d(x, y)\,\mathrm{d}s \tag{4}$$

Assume a rectangular tree of 14-m height, 11-m width, a trunk 8 m away from the receiver, and the receiver 2.4 m above ground (see Fig. 2). The mean total attenuation A_T is then

$$A_T(E) \cong \frac{K\sec(E)}{5.5}\left[\int_{2.5}^{8}(1 - y/14)^2(x - 25)\mathrm{d}x + \int_{8}^{13.5}(1 - y/14)^2(13.5 - x)\mathrm{d}x\right] \tag{5}$$

where $y = 2.4 + x\tan E$. Thus,

$$\begin{aligned} A_T\,(E) &\approx K[3.78\sec E - 5.21\sec E\tan E + 1.937\sec E\tan^2 E] \\ &\cong K[3.726 - 0.0670\,E] \text{ for a least mean square fit} \\ &\cong 24.6 - 0.442\text{ E}. \end{aligned}$$

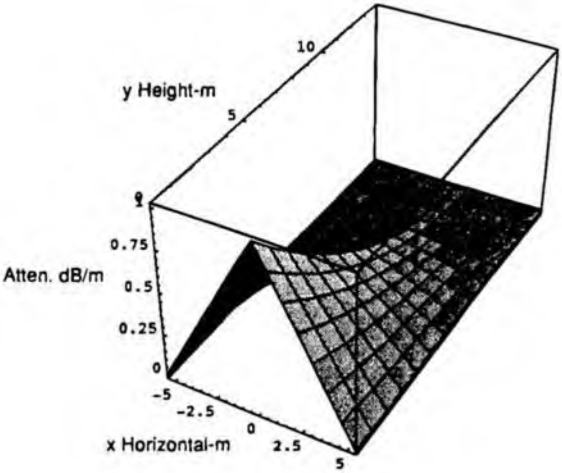


Fig. 3 Simplified mean normalized attenuation density profile of a tree in dB/m that has been adjusted to match measured variation of total attenuation with elevation angle for a callery pear tree of 14-m height and 11-m width at the base. The height is the height above the ground level. The zero point on the x width scale represents the trunk of the tree.

The calculated value of $A_T(E)$ vs elevation angle is shown in Fig. 4 as the dashed curve where K has been selected for a good match to measured data. Also shown in the figure are the measured points from Goldhirsh and Vogel[1] (which have been scaled up to 1.525 from 870 MHz by a factor of 1.345). The solid curve represents the Goldhirsh and Vogel empirical model:

$$A_T(E) = 25.8 - 0.47\ E \qquad (6)$$

which is rather close, as can be seen, to the measured points and the linear best fit result using the geometric model of Fig. 2. We can conclude that the attenuation density is clearly not uniform with position, and the empirical model of Fig. 2, at least for this example, gives a good match for the limited range of elevation data available (between 12 and 30 deg).

Other data have been gathered for various types of trees, as shown in Table 1. The table shows measured tree attenuation for various types of trees at 870 MHz. As discussed earlier, in this frequency region, tree foliage attenuation varies approximately as the square root of frequency; i.e., $\sqrt{f_{L1}/870\ \mathrm{MHz}}$ and results are scaled to 1.575 GHz in terms of increased dB/m attenuation. Trees with full foliage have approximately 1.35 times the attenuation in dB/m compared to attenuation for deciduous trees bare of foliage. Note that some trees have as much as 3–4 dB/m of attenuation at 1.5 GHz. Elevation angle information was not available for these measurements. Note also the substantial difference in two attenuation measurements made for pin oak. The measurements were made at two different locations. Thus, we should expect a wide variation in the data from tree to tree, and the data are primarily useful in a statistical sense.

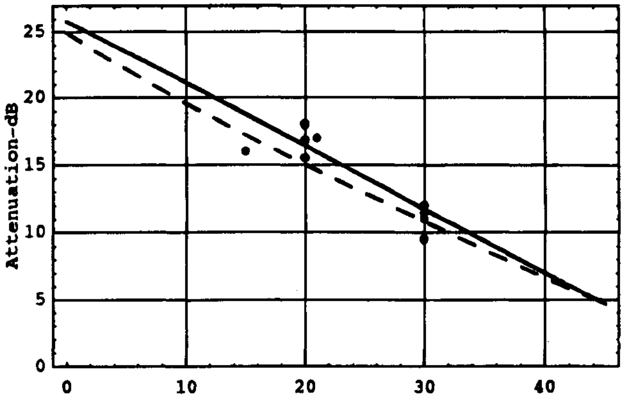


Fig. 4 Measured and approximate attenuation scaled to 1.575 GHz for a callery pear tree of 14-m in height with a base width of 11 m corresponding to Fig. 2. The measured points are shown, the solid curve is the empirical model from Ref. 3, and the dashed curve is the resulting curve from the simple physical model of Fig. 2.

Next, consider the attenuation through a grove of trees, as shown in Fig. 5 where the user–satellite path travels a distance d through the grove. The CCIR[5] has developed an approximate model for mobile user path loss vs the ray foliage penetration distance d of Fig. 5. The CCIR modified exponential decay model gives a path loss $L = \alpha d$ where $\alpha = 1.33 f^{0.284} d^{-0.412}$ dB/m for $14 < d < 400$ m and $\alpha = 0.45 f^{0.284}$ for $0 < d < 14m$ and frequency f in GHz. This model* gives less than a linear dependence over d, which compensates for the Fresnel diffraction effect that gives a decreased attenuation effect with larger user–tree grove distances.[5,6]

A. Foliage Attenuation—Mobile User

Next consider a user who is mobile rather than stationary; i.e., the user is moving rapidly past trees and open spaces alongside the roadway. If these trees and open spaces pass by relatively rapidly compared to the GPS receiver closed loop bandwidth and databit interval of 20 ms = 1/50 bps, then the mean attenuation statistic is more useful than the attenuation of each individual tree.

An empirical model[7] for the fade depth attenuation in dB at 1.5 GHz vs elevation angle E for driving along forested roadways in Maryland is given by the following:

$$F(P, E) = -[a_0 + a_1 E + a_2 E^2]\, ln\, P + bE + c \text{ dB} \tag{7}$$

where P is the percentage of time this fade depth F is exceeded (P = 1–20%); $a_0 = 3.44$; $a_1 = 0.0975$; $a_2 = -0.002$; $b = -0.443$, $c = 34.76$, and E is the elevation angle in degrees to the satellite. A graph of this empirical relationship over the region 20 deg $< E <$ 60 deg for $P = 1\%$, and $P = 10\%$ is shown in Fig. 6.

*Note that this CCIR model does not deal with specifics such as the type of foliage, or the distance from user to the grove of trees.

Table 1 Single tree attenuations at f = 870 MHz and attenuation coefficient scaled dB/m for 1.5 GHz (scale factor of 1.35)[7]

	Attenuation dB—870 MHz		Attenuation coefficient dB/m—870 MHz		Attenuation coefficient dB/m—1.575 GHz	
Tree type	Largest	Average	Largest	Average	Largest	Average
Burr oak	13.9	11.1	1.0	0.8	1.3	1.1
Callery pear	18.4	10.6	1.7	1.0	2.2	1.3
Holly	19.9	12.1	2.3	1.2	3.0	1.6
Norway maple	10.8	10.0	3.5	3.2	4.6	4.2
Pin oak	8.4	6.3	0.85	0.6	1.1	0.8
Pin oak	18.4	13.1	1.85	1.3	2.4	1.7
Pine cone	17.2	15.4	1.3	1.1	1.7	1.5
Sassafras	16.1	9.8	3.2	1.9	4.2	2.5
Scotch pine	7.7	6.6	0.9	0.7	1.2	0.9
White pine	12.1	10.6	1.5	1.2	2.0	1.6
Overall average	14.3	10.6	1.8	1.3	2.4	1.7

The empirical Eq. (7) for fade depth statistics can be rewritten in terms of the probability of a fade level being exceeded as a function of fade level for various elevation angles. This probability is as follows:

$$P(F, E) = \exp[(-17380 + 500\,F + 221.5E)/(-1720 - 48.75E + E^2)] \qquad (8)$$

for $1 < P < 20\%$ and 20 deg $< E >$ 60 deg. The probability of fade exceedance is shown in Fig. 7 for elevation angles of 20, 30, 40, 50, and 60 deg along these forested roadways in Maryland.

B. Probability Distribution Models for Foliage Attenuation—Mobile User

The probability distribution for fading along roadways has one of two different types of models, shadowed or unshadowed. For the unshadowed model, we

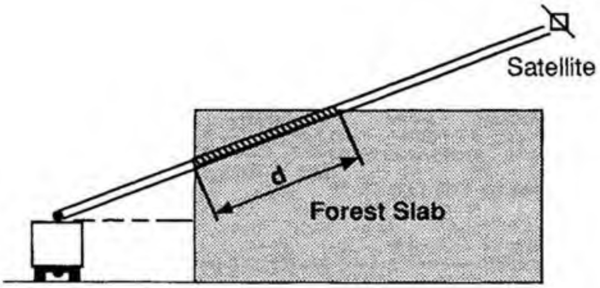


Fig. 5 Model of forest next to roadway. The vegetative path length is d meters.

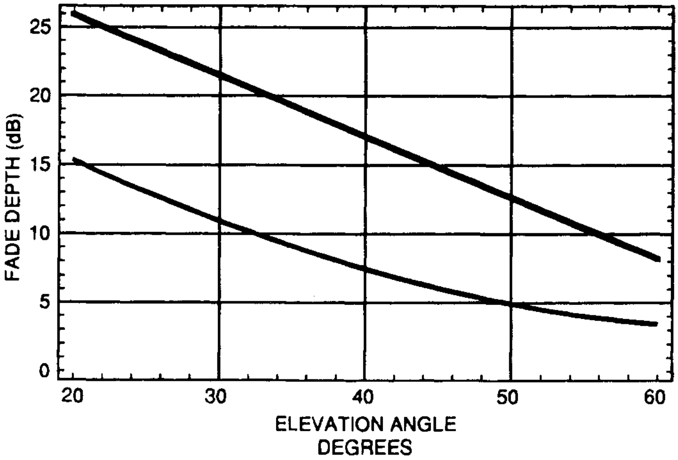


Fig. 6 Fade depth statistics in dB vs elevation angle *E* in degrees for *L*-band satellite to mobile users along a forested roadway. The two curves are for 1% (upper heavy curve) and 10% (light curve) probability that the fade is no more than the value shown.

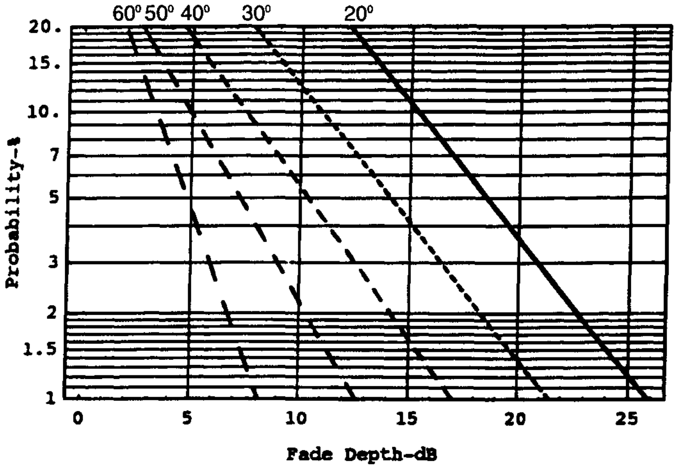


Fig. 7 Probability of fade depth being exceeded vs fade depth for various elevation angles: 20 deg, 30 deg, 40 deg, 50 deg, 60 deg. The solid curve is for an elevation angle of 20 deg, and the curves to the left are for progressively higher elevation angles in 10-deg increments. These empirical data correspond to measurements at *L*-band taken along forested roadways in Maryland. (data from Ref. 7)

assume that the received signal energy consists of the sum of 1) direct ray (constant amplitude); and 2) Rayleigh scattered signals. The received signal then has the following form:

$$
\begin{aligned}
r &= a\cos(\omega_o t + \phi) + \omega\cos(\omega_o t + \theta)\\
&= x_c \cos\omega_o t + x_s \sin\omega_o t
\end{aligned}
\tag{9}
$$

where ω has a Rayleigh distribution with mean square value σ^2 and a is the direct ray signal amplitude. The sum of these two has a Nakagami–Rice distribution for the envelope $z = \sqrt{x_c^2 + x_s^2}$ of the in-phase and quadrature signal components. The probability density of the unshadowed envelope is then as follows:

$$
p_u(z) = \frac{z}{\sigma^2}\exp\left[\frac{-z^2 + a^2}{2\sigma^2}\right] I_o\left(\frac{za}{\sigma}\right) \tag{10}
$$

where the normalized line-of-sight signal power is $P_s = a^2/2$, and the power in the Rayleigh scattered signal is $P_R = \sigma^2$, and $I_o(\ )$ is the modified Bessel function of zero order. If the direct line-of-sight a^2 term decreases to zero, we are left with the Rayleigh signal.

In the shadowed environment, we use a distribution proposed by Loo[2] in which the conditional probability density of the total received signal for a given direct signal level a; namely $p(z|a)$, is given by the Nakagami–Rice distribution, and the probability density of the a direct ray $p(a)$ component is lognormal. The lognormal density is given by the following:

$$
p(a) = \frac{1}{\sigma_o a\sqrt{2\pi}}\exp\left[\frac{-(\ell n\ a - \mu)^2}{2\sigma_o^2}\right] \tag{11}
$$

as shown in Fig. 8. In essence, the lognormal distribution is the same as a Gaussian

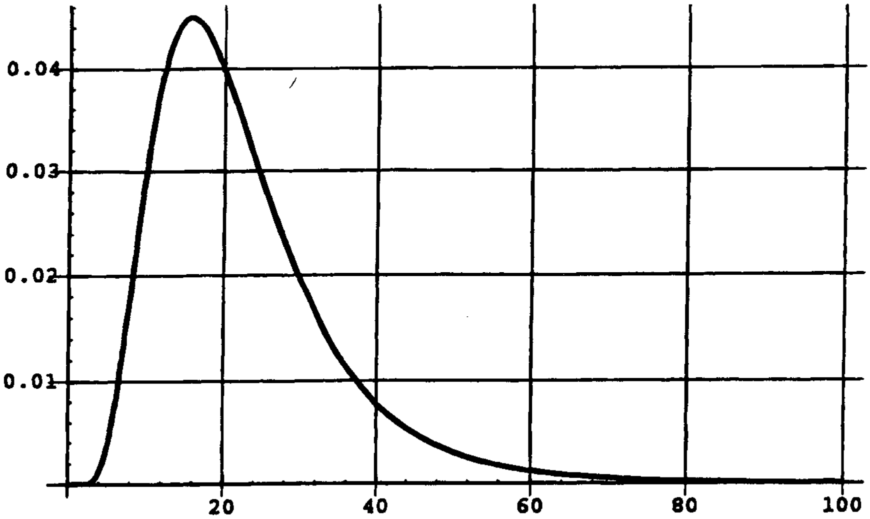


Fig. 8 Probability density functions of the lognormal distribution function for mean $\mu = 3$, and $\sigma_o = 0.5$.

density function where the variable is attenuation or signal level expressed in dB (except for a scale factor to account for the difference between $\log_{10}$ and ℓn).

The shadowed density function formulated by Loo[2] is then the convolution of the Rice and lognormal densities:

$$p_s(z) = \int p(z \mid a)p(a)\,\mathrm{d}a \tag{12}$$

Barts and Stutzman[6] have taken this problem a step further by assigning a probability S that the user is shadowed in the course of travel down a road. During the shadowed condition, the probability density is $p_s(z)$. The signal level has an unshadowed probability density $p_u(z)$ for the remaining fraction $1 - S$ of the time. The composite signal probability density is then as follows:

$$p_c(z) = Sp_s(z) + (1 - S)p_u(z) \tag{13}$$

The coefficients of these probability densities can be fit to measured data to give reasonably good models for the measurements.

The Loo probability density from Eq. (12) is as follows:

$$p_s(z) = \frac{2z}{\alpha\sigma\sqrt{2\pi}} \int_0^\infty \frac{1}{a} I_o\left(\frac{2za}{\alpha}\right) \exp\left[\frac{-(ln\ a - \mu)^2}{2\sigma^2} - \frac{a^2 + z^2}{\sigma}\right] \mathrm{d}a \tag{14}$$

where α is the Rayleigh parameter; and μ, σ are the lognormal mean and sigma. This density is rather cumbersome to use. Barts and Stutzman[6] have suggested an approximation to the Loo density for the envelope of the shadowed signal. Their power–law approximation to the shadowed propagation fade distribution is then as follows:

$$P_s(F) \cong \left(\frac{50 - F}{V_1}\right)^{V_2} \tag{15}$$

where

$$V_1 = -0.275\,\overline{K} + 0.723\mu + 0.336\sigma + 56.979$$

$$V_2 = [-0.006\,\overline{K} - 0.008\mu + 0.013\sigma + 0.121]^{-1}$$

where in the expression all parameters $\overline{K}$, μ, σ are expressed in dB, and $\underline{P_s(F)}$ is the probability that the fade F in dB is exceeded. We have defined $\overline{K} = 10 \log(1/\alpha)$ dB.

A similar approximation was given for the unshadowed Ricean distribution:

$$p_u(z) = p_{\mathrm{Rice}}(z) = \frac{2z}{\beta} \exp\left[\frac{-(z^2 + a^2)}{\beta}\right] I_o\left(\frac{2za}{\beta}\right) \tag{16}$$

where B is the mean square value of the Ricean distribution of r, and C is the amplitude of the direct component. Define $K = 10 \log(a^2/\beta)$ dB. The approximation of Barts and Stutzman[6] is as follows:

$$P_u(F) \cong \exp\left[\frac{-(F + u_1)}{u_2}\right] \tag{17}$$

where

$$u_1 = 0.01K^2 - 0.378K + 3.98$$

$$u_2 = 331.35\ K^{-2.29}$$

and $P_u(F)$ is the probability or fraction of time that a given fade depth F in dB is exceeded.

III. Measured Models—Satellite Attenuation Data

Fade statistics for mobile users in suburban, rural/forested, rural/farmland, mountainous terrain, and tree-lined roads have been measured by a number of investigators. A set of data taken at the moderately low elevation angle of 19 deg is shown for a variety of environments in Fig. 9. Note again that for GPS applications, we generally want access to satellites at low elevation angles for low GDOP, as well as access to a satellite at zenith. As can be seen, fades of 8 dB or more are possible for a reasonable fraction (≈10%) of the time in suburban areas. Fades of 20 dB or more are encountered 1% of the time for suburban and rural forested regions at this low 19-deg elevation angle. It should also be pointed out that there can be a rapid spatial fluctuation in the fading. The deeper fades generally have considerable variation (≈5 dB) for movements of the receiver as small as 2 m for forested roadways.

At higher elevation angles where line-of-sight propagation (unshadowed) is maintained, the attenuation (caused by multipath scattering) is significantly less.

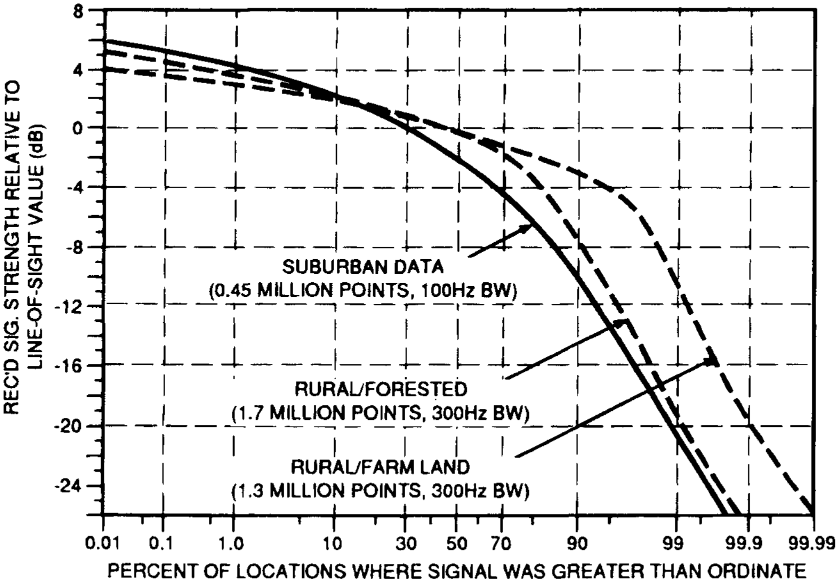
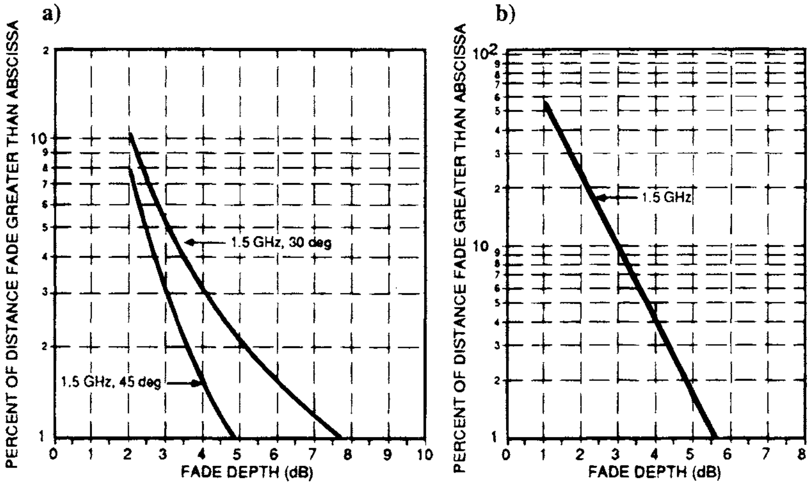


Fig. 9 Measurements of *L*-band propagation statistics for mobile users on land for various environments at an elevation angle of 19°. These plots are based on data from Kent.[8]



Best power curve fit cumulative fade distributions for line-of-sight configurations in mountainous terrain. The cumulative fade distribution probability P for this power law approximation is of the form

$$P = aF^{-b}$$

where F is the fade in dB and where a, b are

	E	
	30°	45°
a	33.19	39.95
b	1.710	2.321

Best exponential fit cumulative fade distributions for line-of-sight configurations on tree-lined roads. The best exponential fit is of the form

$$P = u \exp(-vF)$$

where

$u = 125.6$
$v = 1.116$

are the best u, v values for moderately high elevation angles when averaged over measurements at E = 30°, 45°, 60°

Fig. 10 Unshadowed line-of-sight measurements of *L*-band fade distributions for mobile users on land for various environments and elevation angles. Multipath reflections cause the fading: a) line-of-sight model for mountainous terrain; b) line-of-sight for tree-lined roads. These plots are based on data from Goldhirsh and Vogel.[14]

Figure 10 shows best fit empirical cures at the GPS L_1 frequency 1.5 GHz. The 1% fade depth was about 5 dB for either tree-lined roads or mountainous terrain (with an elevation angle of 45 deg).

A. Measured Fading for Tree-lined Roads—Mobile Users

Numerous measurements have been made worldwide on fading depth as a function of elevation angle as a vehicle is driven around various tree-lined roads. Figure 11 shows a summary of a selected set of these measurements. These measurements include the following:

1) Bundrock and Harvey,[9] Melbourne, Australia—1.55 GHz tree-lined road with 85% tree incidence

2) Butterworth and Mott,[10] Ottawa, Canada—1.5 GHz, 19 deg elevation angle, rural, forested, hilly terrain with immature timber with occasional cleared areas, two-lane road with good shoulders

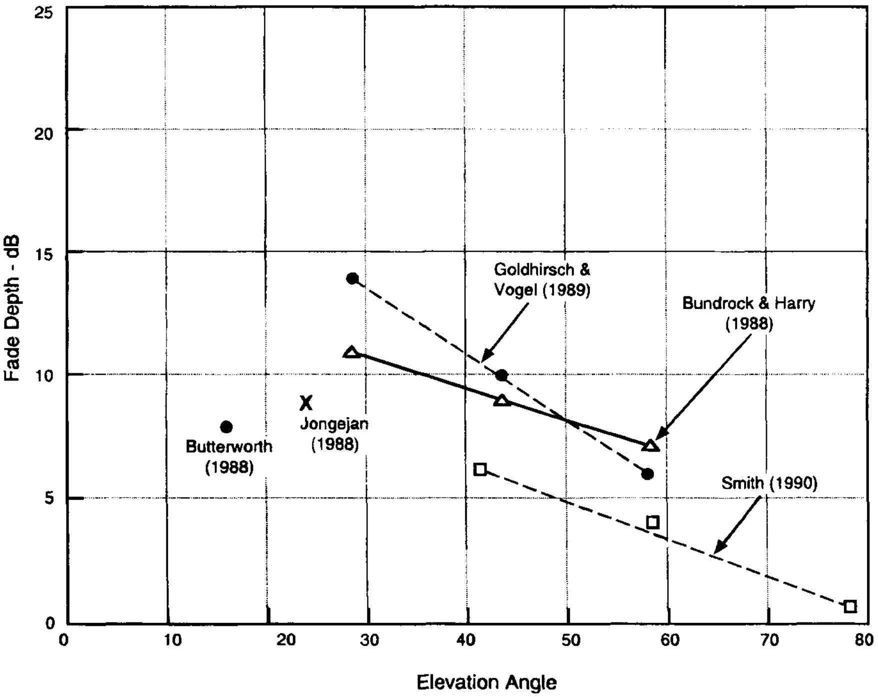


Fig. 11 Summary of various measurements of fading depth for tree-lined roads for 10% probability of the fade level being exceeded.

3) Jongejans et al.,[11] Belgium, 1.5 GHz—hilly Ardennes with roadside lined with bare trees in winter

4) Smith et al.,[12] England—rural, tree-shadowed environment with trees with full leaf cover, *L*-band

5) Goldhirsh and Vogel[3]—forested roads in Maryland, two- and four-lane highways

In general, the 1% fade points are substantially higher by as much as a factor of two in the dB fade level; i.e., a 10 dB fade for 10% might increase to 20 dB for the 1% mark.

These measured data were generally taken with circularly polarized signals of narrow bandwidth. Reference 13 showed that measurements made in a pine forest for vertical and horizontal polarization showed little difference between the two polarizations and concluded that that type of forest at least is approximately an isotropic medium.

References

[1]Goldhirsh, J., and Vogel, W. J., “Propagation Effects for Land Mobile Satellite System: Overview of Experimental and Modeling Results,” NASA Ref. Pub. 1274, Feb. 1992.

[2]Loo, C., “A Statistical Model for a Land Mobile-Satellite Link,” *IEEE Transactions on Vehicular Technology*, Aug. 1985.

[3]Goldhirsh, J., and Vogel, W. J., "Mobile Satellite System Fade Statistics for Shadowing and Multipath from Roadside Trace at UHF and L-Band," *IEEE Transactions on Antenna and Propagation,* April 1989.

[4]Yoshikawa, M., and Kagohara, M., "Propagation Characteristics in Land Mobile Satellite Systems," Thirty-Ninth Annual IEEE Vehicular Technical Conference, May 1989.

[5]CCIR Study Group 1990–1994, "Impact of Propagation Impairments on the Design of LED Mobile Satellite System Providing Personal Communication," CCIR, U.S. WP-8D-14 (Rev. 2), October 1992.

[6]Barts, R. M., and Stutzman, W. L., "Modeling and Simulation of Mobile Satellite Propagation," *IEEE Antenna and Propagation,* April 1992.

[7]Goldhirsh, J., and Vogel, W. J., "Mobile Satellite System Propagation Measurements at L-Band Using MARECS B2," IEEE Trans-Antenna and Propagation, February 1990.

[8]Kent, J. D. B., "A Land Mobile Satellite Data System," International Mobile Satellite Conference, Ottawa, 1990.

[9]Bundrock, A., and Harvey, R., "Propagation Measurements for an Australian Land Mobile-Satellite System," *Proceedings of Mobile Satellite Conference,* 1988.

[10]Butterworth, J. S., and Mott, E. E., "Characterization of Propagation Effects for Land Mobile Satellite Services," International Conference Satellite System for Mobile Communication and Navigation, June 1983.

[11]Jongejans, A. A., et al., "PROSAT—Phase I Report," European Space Agency TR, ISA, STR-216, May 1986.

[12]Smith, H., et al., "Assessment of the Channel Offered by a High Elevation Satellite Orbit to Mobiles in Europe," IEEE Conference on Radio Receiver and Associated Systems, July 1990.

[13]Ulaby, F. T., et al., "Measuring the Propagation Propagation Properties of a Forest Canopy Using Polarimetric Scattermeter," *IEEE Transactions on Antenna and Propagation,* Feb., 1990.

[14]Goldhirsh, J., and Vogel, W. J., "An Overview of Results Derived from Mobile-Satellite Propagation Experiments," Inter. Mobile Satellite Conference, Ottawa, 1990.

Chapter 16

Ephemeris and Clock Navigation Message Accuracy

J. F. Zumberge* and W. I. Bertiger†
Jet Propulsion Laboratory, California Institute of Technology,
Pasadena, California 91109

IN this chapter, we discuss the accuracy of the ephemeris and clock corrections contained in the Global Positioning System (GPS) navigation message. We first provide a brief description of how the Control Segment generates these quantities. Next, we compare them with results from precise (non-real-time) solutions of satellite parameters derived from the simultaneous analysis of data from a globally distributed network of GPS receivers. Finally, we cast these accuracies into the form of a user equivalent range error.

I. Control Segment Generation of Predicted Ephemerides and Clock Corrections

One of the primary purposes of the Control Segment (Chapter 10, this volume) is to generate predicted satellite ephemerides and clock corrections, which are regularly uploaded to the satellites. The predictions are then included as part of the 50-b/s 1,500-bit navigation message (Chapter 4, this volume) that modulates the transmitted GPS signal. Ground receivers then use the predictions for real-time estimates of satellite coordinates and clock corrections.

Data used for the predictions are acquired from receivers situated at precisely known locations in Hawaii, Colorado, Ascension Island in the Atlantic Ocean, Diego Garcia in the northern Indian Ocean, and Kwajalein in the western Pacific. The distribution in longitude of these sites (Table 1) is reasonably uniform, allowing continuous tracking of all GPS spacecraft. The sites at Ascension, Diego Garcia, and Kwajalein are capable of transmitting computed navigation message updates to the satellites. Receivers at all stations use cesium oscillators for time stability, and measure dual-frequency phase and pseudorange. Meteorological data are acquired at each station and used to aid in estimation of troposphere

*Member Technical Staff, Satellite Geodesy and Geodynamics Systems Group, Tracking Systems and Applications Section.

†Member Technical Staff, Earth Orbiting Systems Group, Tracking Systems and Applications Section.

Table 1 Tracking stations used by the Control Segment and approximate locations

Site	Latitude	Longitude
Hawaii	21°N	158°W
Colorado Springs[a]	39°N	105°W
Ascension Island[b]	8°S	14°W
Diego Garcia[b]	7°S	72°E
Kwajalein[b]	9°N	168°E

[a]Master Control Station.
[b]Can transmit to GPS satellites.

delay. All data are regularly transmitted to the Master Control Station in Colorado Springs.

Only the P-code pseudorange measurements are used as data in the parameter estimation scheme, which is based on a Kalman filter. Estimated satellite parameters include epoch-state position and velocity, solar radiation pressure coefficients, clock bias, drift, and drift rate. Station parameters include similar clock quantities and troposheric delay. The terrestrial coordinate system and gravity field are 1984 World Geodetic System (WGS-84). The reference time is an average of monitor station clocks and a subset of GPS clocks.

Data going back 4 weeks are used to estimate reference satellite trajectories, which are then used to propagate satellite positions and clock corrections into the future. The first 28 h of prediction are divided into overlapping 4-h fit intervals separated by 1 h. The fit results for each such interval are cast in the format of the navigation message (through a fitting procedure), and are uploaded into the satellites once a day, more frequently if required to meet a 10-m user-equivalent range error specification. The daily upload is based on a data window that closed 45 min prior to the upload.[1–3]

Given the daily upload, the satellite broadcasts satellite positions and clock corrections contained in the appropriate 4-h interval. Although predictions beyond 28 h are also uploaded, they are normally not used, because the next day's upload overwrites them with results derived from more current data.

II. Accuracy of the Navigation Message

This section assesses the accuracy of the information broadcast in the navigation message. The "truth cases" to which the navigation messages are compared are daily GPS solutions, from the Jet Propulsion Laboratory (JPL), of satellite positions and clock corrections. First we describe the daily JPL solutions, including estimates of their accuracies. Next, based on the period 1993 July 4–Oct. 22, we compare these daily solutions with their counterparts from the navigation message. Of course, it must be remembered that one of the key differences between the GPS Control Segment solutions and the truth model is that the GPS solutions are predictions of the future based on past data, whereas the "truth" solutions are based on after-the-fact, postfit smoothed estimates.

A. Global Network GPS Analysis at the Jet Propulsion Laboratory

Since June 1992, analysts at the JPL have regularly reduced GPS data from a globally distributed network of 20–40 precision P-code GPS receivers using the GIPSY/OASIS-II software.[4,5] Shown in Fig. 1 are locations of the sites as of fall 1993. In addition to dense coverage in North America and Europe, there is also reasonable coverage elsewhere, including eight sites in the Southern hemisphere.

Receivers at these sites make measurements of the carrier phase and pseudorange observables on both the L_1 and L_2 bands from GPS satellites, at 30-s data intervals. Data are analyzed daily in 30-h batches, centered on GPS noon. The 6-h overlap centered at each GPS midnight allows for consistency checks between solutions from adjacent days. Prior to parameter estimation, data are edited using the TurboEdit alogrithm[6] and decimated to a 10-min interval.

The model used in the analysis corrects for ionospheric delay (through the formation of the ionosphere-free linear combination of phase and pseudorange observables), tropospheric delay (by stochastic estimation of the wet component at each receiver site), transmitter and receiver phase center offsets, Earth orientation (through explicit estimation of pole position and length of day), solid Earth tides, and relativistic effects. Transmitter and receiver clock corrections are estimated as independent parameters at each sample time. [In the case of selective availability (SA)-affected transmitters, this accounts for the dithering of GPS clocks.] The reference clock is a hydrogen maser driving one of the receivers. Satellite parameters include epoch-state position and velocity, and solar radiation pressure; the latter is estimated stochastically. The Earth-fixed reference frame is defined by adopting fixed locations for eight of the receivers, as specified in the international terrestrial reference frame, ITRF-91.[7,8] The carrier phase biases are estimated as piecewise-constant, real-valued parameters.

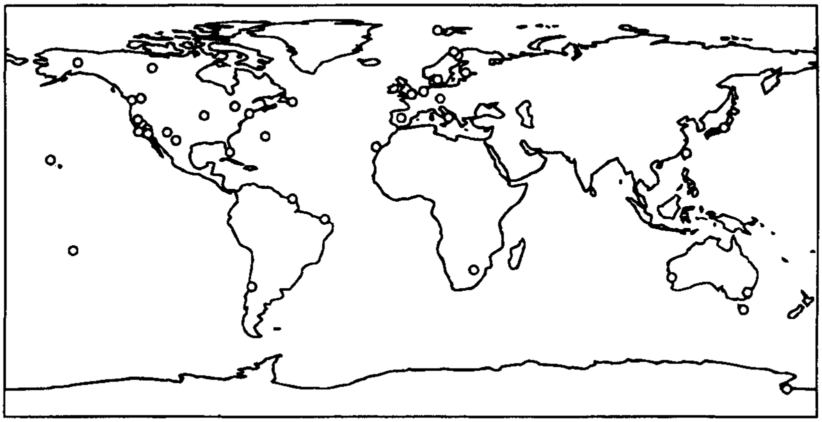


Fig. 1 Global distribution of GPS tracking receivers in the International GPS Service for Geodynamics, fall 1993. In addition to dense coverage in North America and Europe, there is also reasonable coverage elsewhere, including eight sites in the southern hemisphere.

B. Accuracy of the Precise Solution

One assessment of orbit quality from the precise solutions can be made by looking at the continuity of results from adjacent days. For example, the solution for Oct. 7 uses data from 2100 h on Oct. 6 to 0300 h on Oct. 8. Similarly, the solution for Oct. 6 uses data from 2100 h on Oct. 5 to 0300 h on Oct. 7. The difference between these solutions in the position of satellite PRN 12 within ± 3 h of midnight between Oct. 6 and 7 is shown in Fig. 2. The rms variation over the 6-h period is 12 cm, 22 cm, and 25 cm for the radial, cross-track, and along-track components, respectively.

A number of other groups estimate satellite parameters from the same data, using independent software, thus allowing a separate assessment of orbit quality. Shown in Fig. 3 is the comparison of the JPL's orbit solution for PRN12 on Oct. 7, 1993 with that determined by the Center for Orbit Determination in Europe (CODE) at the University of Berne, Switzerland.[9] The vertical scale is the same as that in Fig. 2. The rms difference over the day between the solutions is 7 cm for the radial component, 5 cm for the cross-track, and 9 cm for the along-track. For other days and satellites, this agreement is generally within 20 cm for all components. Comparisons with results from other analysis centers show comparable agreement. To summarize, the precise orbit solutions are typically accurate to 5–30 cm rms, depending on the component and other factors.

A comparison similar to that shown in Fig. 2 is shown for a transmitter clock in Fig. 4. Plotted there is the difference between the Oct. 6 and Oct. 7 solution for the PRN13 clock. The rms difference over the 6-h period is 0.22 ns, of which a portion is attributable to a 0.18-ns bias. This few-tenths-ns rms difference is typical for other satellites and days.

The reference for the precise clock solutions is the maser-based receiver at Algonquin Park, Canada, maintained by the Geodetic Survey Division of Canada's Department of Energy, Mines and Resources. This clock is adjusted periodically,

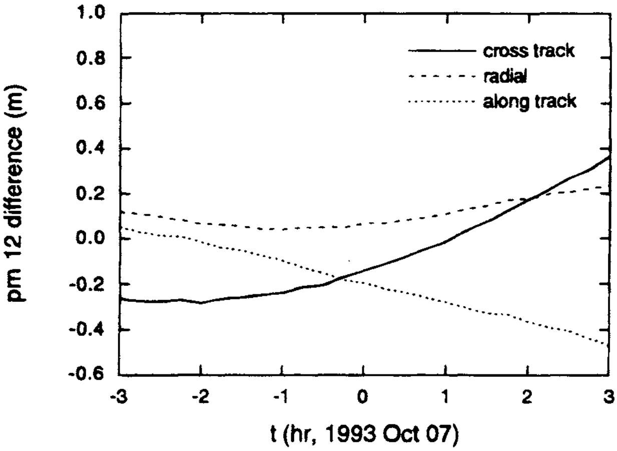


Fig. 2 Difference near the midnight boundary between the JPL's precise solutions of Oct. 6, 1993 and Oct. 7. The rms differences are 0.12 m, 0.22 m, and 0.25 m in the radial, cross-track, and along-track components.

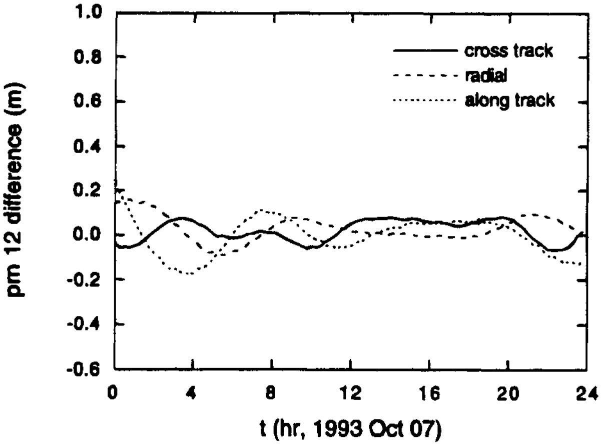


Fig. 3 Comparison of the precise solution from the JPL with that from the CODE (University of Berne), for PRN 12 on Oct. 7, 1993. The rms values over the day (which include both the bias and the variation) are 0.07 m for the radial component, 0.05 m for the cross-track, and 0.09 m for the along-track. The agreement between the JPL and CODE for this satellite and day is somewhat better than typical. However, for other satellites and days the agreement is, with few exceptions, at least as good as a few tens of centimeters.

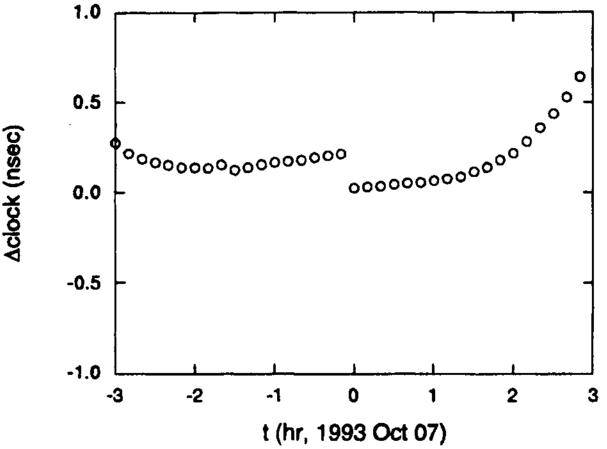


Fig. 4 Difference in the clock solution for PRN 13 from the Oct. 6, 1993 solution and the Oct. 7, 1993 solution. The rms value over the 6-h overlap is 0.22 ns (of which a portion is attributable to a 0.18-ns bias).

and is believed accurate with respect to GPS time to within a few hundred ns, with drift magnitudes of no more than a few tens of ns per day.[10] Of course, biases and drifts in the reference clock will be masked in a comparison such as Fig. 4, but would appear in a comparison between the navigation message and the precise solutions.

C. Comparison of Precise Orbits with Broadcast Ephemerides

Fig. 5 shows the position difference between the JPL precise solution and the broadcast orbit for satellite PRN 12 on Oct. 7, 1993. The interval between updated navigation messages typically is one hour. The satellite and day are the same as in Figs. 2 and 3, although the vertical scale is 10 times larger. The rms values over the day are 0.56 m for the radial component, 1.67 m for the cross-track, and 2.67 m for the along-track.

Similar calculations have been made for all satellites and days over the period July 1, 1993 through Oct. 22, 1993 (a total of 2490 satellite days). For the given satellite and day, the rms difference over the day between the broadcast ephemeris and the precise solution is computed, for each of the three components. The results are summarized in Fig. 6.

Figure 6 contains three histograms, one for each of the position difference components. The median values in the above distributions are 1.3 m for the radial component, 3.6 for the cross-track, and 4.7 m for the along-track. Thus, half of the satellites and days over the ~4-month period had a daily rms agreement between the navigation message and the precise solution of less than 1.3 m in the radial component.

The reference frame ITRF-91 used for the precise solutions differs from the 1984 World Geodetic System (WGS-84) used by the broadcast ephemeris. To test how much this reference-frame difference contributes to the observed

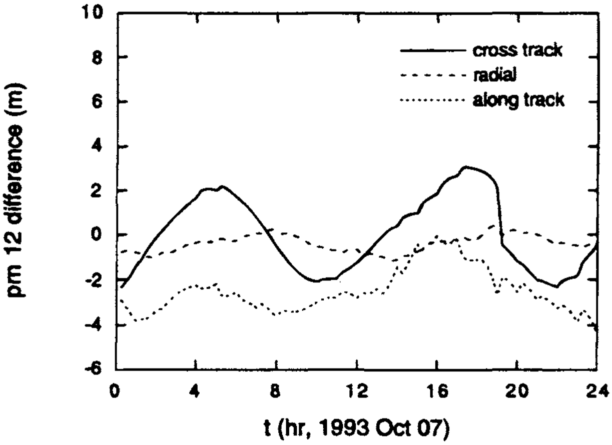


Fig. 5 Comparison of the GPS broadcast ephemeris in the navigation message with the JPL's precise solution, for PRN 12 on Oct. 7, 1993. The rms values over the day are 0.56 m for the radial component, 1.67 m for the cross-track, and 2.67 m for the along-track. The vertical scale is 10× that of Figs. 2 and 3.

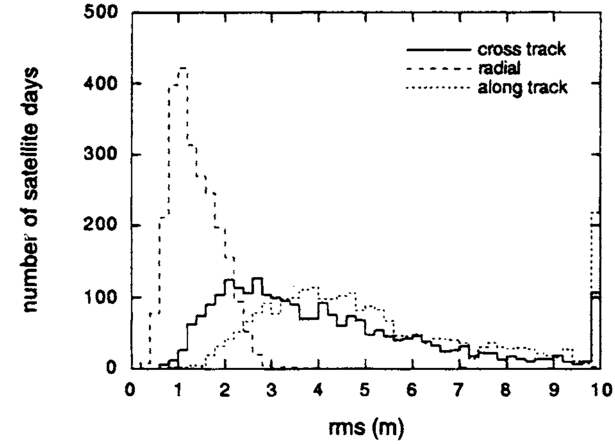


Fig. 6 Comparison of the GPS broadcast ephemeris with precise orbital solutions for the period July 4, 1993 through Oct. 22, 1993. An "event" in one of the three histograms corresponds to a single satellite on a single day (as in Fig. 5). For the given satellite and day, the rms difference over the day between the broadcast ephemeris and the precise solution is computed, for each of the three components. The median values in the above distributions are 1.3 m for the radial component, 3.6 m for the cross-track, and 4.7 m for the along-track. If a daily seven-parameter transformation is applied to align the navigation message reference frame with that of the precise solution, the medians are marginally reduced to 1.2 m, 3.2 m, and 4.5 m.

difference in ephemerides, consider the 7-parameter transformation defined by $x' = (1 + \epsilon)x + T + \Theta x$, where T is a translation vector, ϵ a scale factor, and

$$\Theta \equiv \begin{pmatrix} 0 & \theta_z & \theta_y \\ \theta_z & 0 & -\theta_x \\ -\theta_y & \theta_x & 0 \end{pmatrix}$$

is a rotation matrix. The difference between x and x' can be thought of as arising from 1) a shift or translation T; 2) a change in orientation, characterized by Θ; and 3) an overall expansion or contraction, characterized by ϵ.

The calculations that resulted in Fig. 6 were repeated, but each day a transformation was applied to all coordinates in the broadcast message. The parameters were chosen to minimize $\Sigma_{pct}\,[\Delta^2_{pct}]$ where $\Delta_{pct} \equiv X_{pct} - x'_{pct}$ is the difference at time t between the precise orbit (X_{pct}) and the transformed broadcast orbit (x'_{pct}) for PRN p and Cartesian component c.

The median values of the daily rms differences in the radial, cross-track, and along-track components are reduced to 1.2 m, 3.2 m, and 4.5 m, respectively, from the values corresponding to Fig. 6. The values of the transformation parameters are given in Table 2, and Fig. 7 shows the daily values of the scale factor.

There are two known effects which would contribute to the scale factor parameter ϵ. The first is that the precise orbits refer to the spacecraft center

Table 2 Parameters in the transformation from the broadcast ephemeris reference frame to that of the precise orbits, based on 120 daily transformations[a]

Parameter	Average	Standard deviation
T_x (cm)	-4.0 ± 2.3	24.7
T_y (cm)	6.3 ± 1.8	19.6
T_z (cm)	-4.6 ± 3.3	36.3
ϵ (ppb)	-12.7 ± 0.2	2.2
θ_x (nrad)	-7.9 ± 3.2	34.9
θ_y (nrad)	-6.8 ± 3.3	35.9
θ_z (nrad)	-125.5 ± 4.3	47.2

[a] The uncertainties in the average values are based on observed daily fluctuations, shown in the third column, divided by $\sqrt{120}$. The most significant parameters are the -12.7×10^{-9} scale factor and the -125.5-nrad rotation around the z-axis.

of mass, whereas the broadcast orbits refer to the spacecraft antenna phase center. Because these points are separated by 0.9519 m, mostly in the radial direction, we could expect a contribution of about $0.9519/(26.55 \times 10^6) \approx +35.9$ ppb to ϵ in the broadcast-to-precise transformation.

Second, the broadcast orbits use the WGS-84 gravity field, with GM $= 3.986005 \times 10^5$ km^3 s^{-2}, compared with the JGM2 value (Nerem et al.) of 3.986004415×10^5 km^3 s^{-2} used in the precise solutions. Because the radius varies with $(\mathrm{GM})^{1/3}$, we would expect a contribution to ϵ of $-$ 1/3 (3.986005 $-$ 3.986004415)/3.986004415 $\approx -$ 48.9 ppb. (The larger value of WGS-84 would put the satellite out further, requiring a negative value of ϵ to bring it into agreement with the precise orbit.) The sum of these expectations, -13.1 ppb, is remarkably close to the observed value of -12.7 ppb in Table 2.

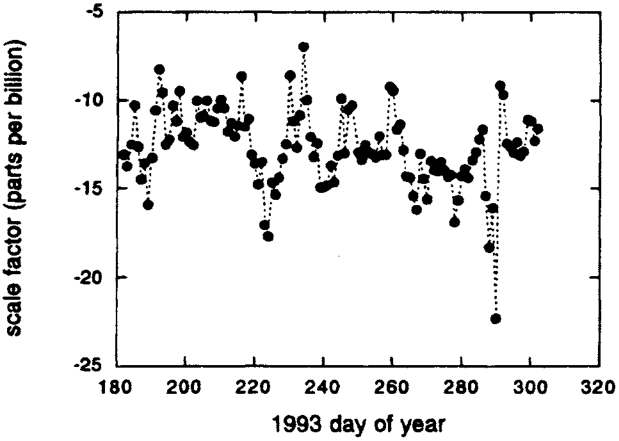


Fig. 7 Scale factor in the transformation from the WGS-84 reference frame to the International Terrestrial Reference Frame. The average value is -12.7 ± 0.2 parts per billion, which corresponds to about -30 cm in the radial direction.

Note that we could use the values in Table 2 to transform WGS-84 coordinates of terrestrial sites to the ITRF-91 reference frame (at epoch midway between July 1 and Oct. 22, 1993), except that ϵ should be taken as $(-12.7 - 35.9)$ ppb $= -48.6$ ppb.

The differences between the precise and broadcast orbits are significantly larger than the accuracies of the precise solutions, as estimated in the previous section. This is not surprising, as the broadcast solutions are, by necessity, the result of an extrapolation in time from hours-old data. The precise solutions, on the other hand, do not have such a real-time constraint. It should also be mentioned that the intentional degradation of broadcast ephemeris quality, one speculated aspect of selectve availability (SA), has not been observed.

D. Comparison of Precise Clocks with Broadcast Clocks

Shown in Fig. 8 are the the clock corrections for PRN 13 on Oct. 7, 1993, as determined by the JPL's precise solution (solid squares) and that from the navigation message (open circles). The interval between points is 30 min. (Points whose estimated uncertainty—"formal error"—exceeds 10 ns in the precise solution are not considered.) The difference between the navigation message and the precise solution is shown in Fig. 9, where the discontinuities in the former are clearly evident.

The GPS satellites on this day can be grouped into two classes depending on the variation over the day in the difference between the precise and broadcast clocks. Figures 9 and 10 contain satellites in the group for which this scatter is of the order of 10 ns. Of the five satellites in this group, three are Block I (PRNs 3, 12, and 13) and two are Block II (PRNs 15 and 20). Note that the time series in Figs. 9 and 10 are reasonably smooth with time.

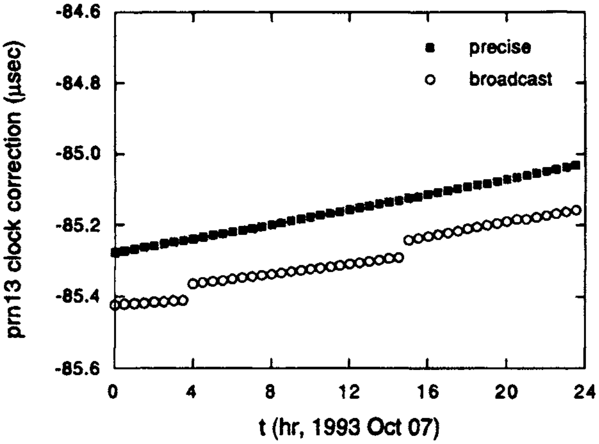


Fig. 8 Clock correction for PRN 13 broadcast in the navigation message (open circles) and determined in JPL's precise solution (solid squares). The precise solution uses a maser-based GPS receiver as its reference. The solid squares describe a slope of about 250 ns/day.

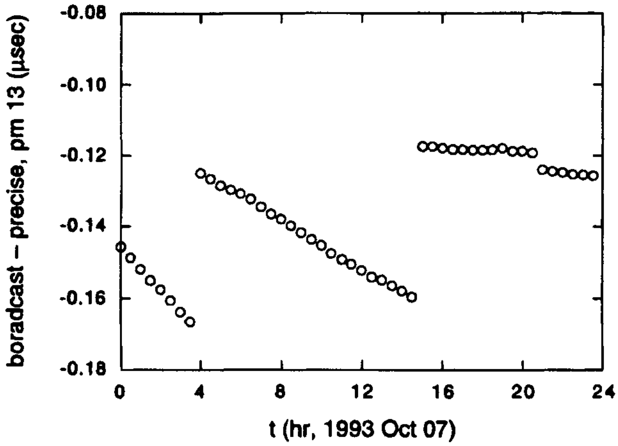


Fig. 9 Difference in the broadcast and precise clock corrections for PRN 13. The discontinuities are attributable to new broadcast messages, as indicated in Fig. 8.

The second class consists of satellites for which the difference is much noisier. All satellites in this group are Block II, and are shown in Fig. 11 as a function of time, together with the histogram that indicates the distribution of the differences. The standard deviation of the distribution is about 80 ns.

The clock dithering component of SA is clearly evident in this second class and absent from the first class. Note that the biases of the distributions in Figs. 9–11 are all about −120 ns, and thus, represent a constant difference between the GPS reference time and that of the precise solution. This could be entirely attributable to the maser-based reference clock used for the latter.

Similar analyses of GPS clocks were made for all satellites and days from July 4, 1993 through Oct. 22, 1993. For each day and satellite, any linear trend in the difference between the precise solution and the broadcast clock over the day was removed. The standard deviation over the day of the detrended difference has been calculated. The distribution of daily standard deviations is shown in Fig. 12. The median of the lower distribution is 4.5 ns, which represents the nonsystematic component of the broadcast clock error for satellites not affected by clock dithering. The median of the upper distribution is 79.9 ns; this quantifies the effect of clock dithering.

A second, systematic component of clock error arises because of differences among satellites each day in the linear trends. (The *average* trend parameters arise from the difference between GPS system time and the time of the Algonquin maser reference; such a difference is not included here because it will have essentially no effect on user position.) Based on the variations in trend parameters among the non-SA satellites over each day, the median value of this systematic component, over all days in the period being studied, has been calculated to be 10.3 ns. The total clock error for non-SA satellites is thus $(4.5^2 + 10.3^2)^{1/2}$ ns ≈ 11.2 ns. For satellites affected by clock dithering it is $(79.9^2 + 10.3^2)^{1/2}$ ns ≈ 80.6 ns.

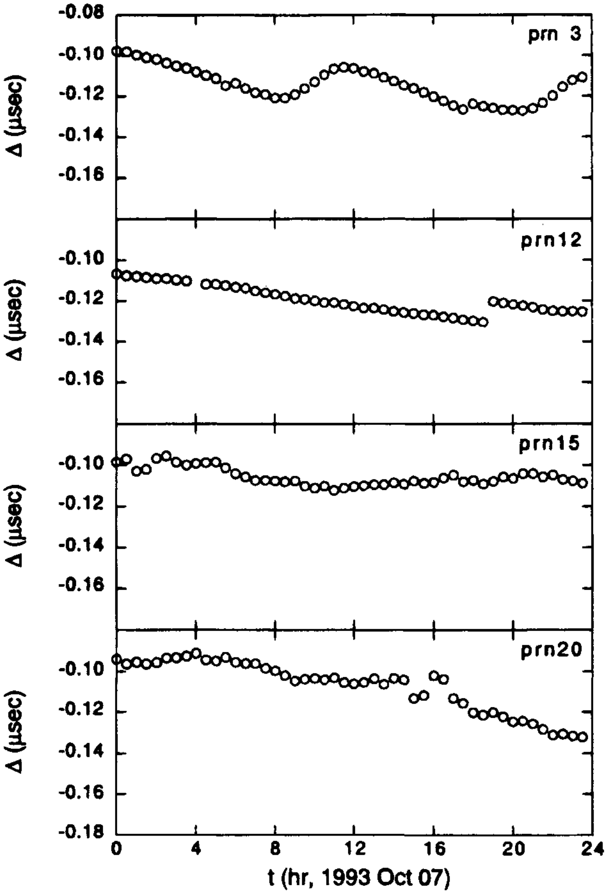


Fig. 10 Difference in the GPS broadcast and precise clock corrections, Δ, for two Block-I spacecraft (PRNs 3 and 12) and two Block-II spacecraft (PRNs 15 and 20). Full scale on each plot is 100 ns. The bias of about −120 ns in all of these (as well as that for PRN 13 in Fig. 9) represents a constant difference between GPS reference time and that of the precise solution.

E. Summary and Discussion

Table 3 compares these results with a prediction[11] of how well the GPS Control Segment would be able to predict clock corrections and GPS ephemerides. The prediction is reasonably consistent with the comparisons between the navigation message and the JPL precise solution, as discussed in this section.

A distribution of differences can be divided into two components: an average (bias) and deviation about that average. Our rms values for orbits include both. Our rms values for orbits include both of these components. Because the transformation from WGS-84 to ITRF will absorb most of the biases, and because the remaining rms values are not reduced much (see Sec. II.C) follow-

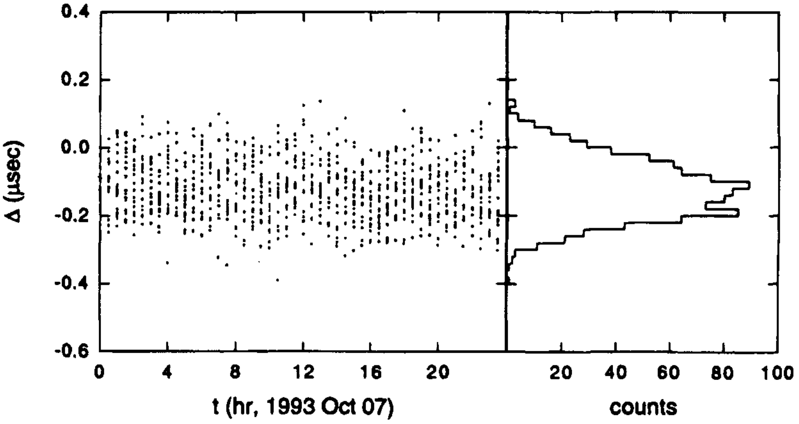


Fig. 11 The effects of clock dithering in SA are shown here, which includes all satellites not shown in Figs. 9 or 10. The distribution of Δ (right) has a standard deviation of about 80 ns and a mean of about −120 ns. (The mean is about the same as those in in Figs. 9 and 10.)

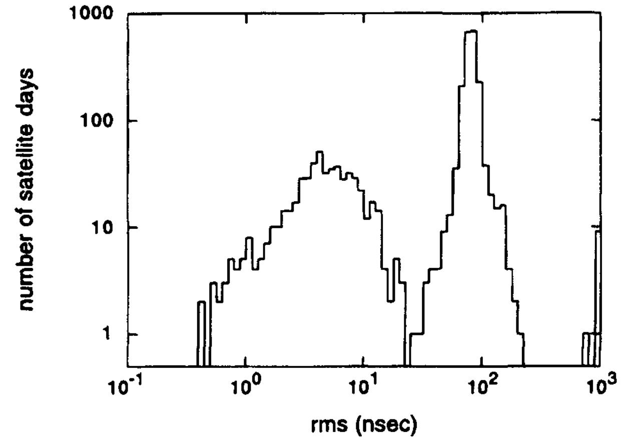


Fig. 12 Comparison of broadcast clock solution with precise solution, for each day and satellite during the period July 4, 1993 through Oct. 22, 1993. The bimodal distribution arises from the effects of clock dithering as part of SA. (There are also a handful of satellites and days when there was rather poor agreement between the precise solution and the broadcast clock.) The median value of the lower distribution (clock dithering presumably not in effect, rms de-trended difference over the day between precise solution and broadcast clock less than 25 ns) is 4.5 ns. The median value of the upper distribution is 79.9 ns.

Table 3 GPS Control Segment performance, predicted and observed

Parameter	Prediction	Observed
Radial (σ_r)	0.8 m	1.2 m
Cross-track ($\sigma_\dagger$)	3.0 m	3.2 m
Along-track (σ_a)	6.3 m	4.5 m
Clock (no SA) (σ_t)	7.7 ns	11.2 ns
Clock (with SA)	—	80.6 ns

ing the transformation, the dominant component is the deviation and not the bias.

The radial uncertainty, just over a meter, is about a factor of 3–4 less than the cross- and along-track uncertainties. This occurs because the range measurement is more sensitive to changes in the radial dimension than to the other dimensions (see Appendix).

We can cast the performances from Table 3 into a user equivalent range error $\sigma_u : \sigma_u^2 = k_r\sigma_r^2 + c^2\sigma_t^2 + k_\rho\rho_{rt}\sigma_r c\sigma_t + k_{\dagger a}(\sigma_\dagger^2 + \sigma_a^2)$. Here, ρ_{rt} is the correlation coefficient between the radial and clock errors; c is the speed of light; and $\sigma_\dagger$, σ_a are the cross- and along-track errors. For a satellite at zenith, an error in the radial component maps directly into σ_u (this is always true for the clock error), while cross- and along-track errors do not affect the user range. We expect nominally, then, that $k_r = 1$, $k_\rho = 2$, and $k_{\dagger a} = 0$. We show in the Appendix that $k_r \approx 0.959$, $k_\rho \approx 1.959$, and $k_{\dagger a} \approx 0.0204$ account for the average geometry in the relationship between satellite and user positions.

It is expected that the correlation between radial and clock errors is small, in which case $|\rho_{rt}| << 1$. For completeness, however, we indicate in Table 4 the contributions to σ_u over the range of possible correlations, that is $\rho_{rt} = 0, \pm 1$.

When SA clock dithering is not in effect, the contributions to σ_u from the ephemeris and clock errors are of the same order, and result in $\sigma_u \approx$ 3–4 m. The more usual circumstance, however, has clock dithering in effect, in which case the dominant contribution to σ_u is from the 80-ns noise in the broadcast clock, resulting in $\sigma_u \approx 24$ m.

Appendix: User Equivalent Range Error

Choose a spherical coordinate system with origin at the center of the Earth. Suppose a satellite has the following Cartesian coordinates:

Table 4 User equivalent range error (meters)

					σ_u		
	σ_r	$c\sigma_t$	$\sqrt{(2\sigma_r c\sigma_t)}$	$\sqrt{[k_{\dagger a}(\sigma_\dagger^2 + \sigma_a^2)]}$	$\rho_{rt} = -1$	$\rho_{rt} = 0$	$\rho_{rt} = +1$
No SA	1.2	1.3	1.8	0.8	0.7	1.9	2.6
With SA	1.2	24.0	7.6	0.8	22.8	24.0	25.1

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} r \sin\Theta \, \cos\Phi \\ r \sin\Theta \, \sin\Phi \\ r \cos\Theta \end{pmatrix}$$

and an observer on the Earth has the following Cartesian coordinates:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r_e \sin\theta \cos\phi \\ r_e \sin\theta \sin\phi \\ r_e \cos\theta \end{pmatrix}$$

where $r_e \approx 6370$ km, and $r \approx 26{,}550$ km. We use θ and Θ to indicate polar angle, and ϕ and Φ to indicate azimuth.

The user range is $u = \sqrt{(X - x)^2 + (Y - y)^2 + (Z - z)^2} = u(r, \Theta, \Phi, r_e, \theta, \phi)$. Choose the orientation of the axes so that 1) the z-axis intersects the satellite ($\Theta = 0$, nominally); and 2) an increase in Θ at $\Phi = 0$ corresponds to the direction of the resultant of the cross- and along-track uncertainties in the satellite position. (We assume that errors in different orbit components are not correlated.) Then, the user equivalent range error is as follows:

$$\sigma_u^2 = c^2\sigma_t^2 + \left(\frac{\partial u}{\partial r}\right)^2 \sigma_r^2 + 2\,\frac{\partial u}{\partial r}\,\rho_{rt}\,\sigma_r c\,\sigma_t + \left(\frac{\partial u}{r\partial\Theta}\right)^2 (\sigma_{\dagger}^2 + \sigma_a^2)$$

(Note that, at $\Theta = \Phi = 0$, we have $\partial u/\partial X = \partial u/r\partial\Theta$.) Differentiations of $u(r,\Theta,\Phi,r_e,\theta,\phi)$ evaluated at $\Theta = 0$ give the following:

$$\frac{\partial u}{\partial r} = \frac{r - r_e \cos\theta}{\sqrt{r^2 + r_e^2 - 2\,r\,r_e \cos\theta}}, \qquad \left(\frac{\partial u}{\partial r}\right)^2 = \frac{(r - r_e \cos\theta)^2}{r^2 + r_e^2 - 2\,r\,r_e \cos\theta}$$

and

$$\left(\frac{\partial u}{r\partial\Theta}\right)^2 = \frac{r_e^2 \sin^2\theta \cos^2\phi}{r^2 + r_e^2 - 2\,r\,r_e \cos\theta}$$

If we assume uniform distribution of positions on the Earth, then we have $p(\theta,\phi)\,d\theta\,d\phi \propto \sin\theta\,d\theta\,d\phi$ as the joint probability distribution of θ and ϕ, for θ less than its maximum value of $\theta_{\max} = \cos^{-1}(r_e/r) \approx 76$ deg. If we average $\partial u/\partial r$, $(\partial u/\partial r)^2$, and $(\partial u/r\partial\Theta)^2$ over $0 \le \theta \le \theta_{\max}$ and $0 \le \phi \le 2\pi$ with this distribution as the weighing function, we obtain $\sigma_u^2 \approx 0.959\,\sigma_r^2 + c^2\,\sigma_t^2 + 1.959\,\rho_{rt}\,\sigma_r\,c\,\sigma_t + 0.0204\,(\sigma_{\dagger}^2 + \sigma_a^2)$.

References

[1]Shank, C., and Smetek, R., personal communication.

[2]Brown, K. R., Jr., "Characterizations of OCS Kalman Filter Errors," *Proceedings of ION GPS-91*, ION, Washington, DC, Sept. 1991.

[3]Bowen, R., Swanson, P. L., Winn, F. B., Rhodus, N. W., and Feess, W. A., "Global Positioning System Operational Control System Accuracies," *Navigation*, Vol. 32, No. 2, 1985.

[4]Webb, F. H., and Zumberge, J. F., (eds.). "*An Introduction to GIPSY/OASIS II*," JPL Course Notes, Boulder, CO, JPL D-11088, July 1993.

[5]Wu, S. C., et al., Topex/Poseidon Project: Global Positioning System (GPS) Precision Orbit Determination (POD) Software Design, JPL D-7275, March 1990.

[6]Blewitt, G., "An Automatic Editing Algorithm for GPS Data," *Geophysical Research Letters*, Vol. 17, No. 3, 1990, pp. 199–202.

[7]Boucher, C., Altamimi, Z., Duhem, L., "*ITRF-91 and Its Associated Velocity Field,*" IERS TN 12, Observatoire de Paris, Oct. 1992.

[8]Blewitt, G., Heflin, M. B., Webb, F. H., Lindqwister, U. J., and Malla, R. P., "Global Coordinates with Centimeter Accuracy in the International Terrestrial Reference Frame using GPS," *Geophysical Research Letters,* Vol. 19, No. 9, 1992, pp. 853–856.

[9]Rothacher, M., Beutler, G., Gurtner, W., Brockmann, E., and Mervart, L., "Bernese GPS Software Version 3.4," Documentation May 1993, Druckerei der Universitaet Bern, Bern, Germany, May 1993.

[10]Kouba, J., and Tetreault, P., "International GPS Service for Geodynamics (IGS)," Natural Resource, Canada, Analysis Reports available every week from the IGS Central Bureau at the JPL (access, for example, with lynx http://igscb.jpl.nasa.gov/).

[11]Rusell, S. S., and Schaibly, J. H., "Control Segment and User Performance," *Global Positioning System, Papers published in Navigation, 1980,* The Institute of Navigation, Washington, DC, 1980.

[12]Wells, D., et al., *Guide to GPS Positioning,* Canadian GPS Associates, Fredericton, New Brunswick, Canada, 1986.

Chapter 17

Selective Availability

Frank van Graas* and Michael S. Braasch†
Ohio University, Athens, Ohio 45701

I. Goals and History

SELECTIVE availability (SA) is the intentional degradation of the Global Positioning System (GPS) signal with the objective to deny full position and velocity accuracy to unauthorized users. Selective availability is part of the Standard Positioning Service (SPS), which was formally implemented on 25 March 1990. Although there have been limited periods of time during which the levels of SA were fairly benign, the GPS policy[1] continues to state the following:

> SPS is planned to provide, on a daily basis, the capability to obtain horizontal positioning accuracy within 100 meters (2 drms, 95 percent probability) and 300 meters (99.99 percent probability), vertical positioning accuracy within 140 meters (95 percent probability), and timing accuracy within 340 ns (95 percent probability).

Clearly, this policy indicates that SA will be active, because the performance of GPS without SA would be in the 20-m range for horizontal positioning accuracy (95%).

SA was not part of the experimental design of GPS. However, initial testing of the Coarse/Acquisition (C/A) code during the 1970s revealed accuracies that were much better than anticipated. The C/A code provided 20–30 m position accuracies rather than the predicted accuracy of no better than 100 m.[2] This prompted the DOD to degrade intentionally the accuracy available to unauthorized users. Initially, the level of SA was set at 500 m (95%), but this was changed to 100 m (95%) in 1983. This level of accuracy was chosen because it is comparable with that provided by an on-airport VHF omnidirectional range (VOR) during the nonprecision approach phase of flight.

II. Implementation

The GPS position solution is obtained by solving a set of four or more pseudorange equations (i = 1 through the number of measurements)

*Associate Professor, Department of Electrical and Computer Engineering.

†Assistant Professor, Department of Electrical and Computer Engineering.

$$\rho_i = \sqrt{(X - X_i)^2 + (Y - Y_i)^2 + (Z - Z_i)^2} + (\Delta t_{\mathrm{RCVR}} - \Delta t_{\mathrm{sv}_i})c \qquad (1)$$

where the unknowns are the receiver position X, Y, Z and the receiver clock with respect to GPS time, Δt_{RCVR}. Propagation delays and other error sources are omitted from Eq. (1). The following parameters are required to solve for the receiver position and clock offset:

ρ_i	measured pseudorange
(X_i, Y_i, Z_i)	satellite position for satellite number i
Δt_{sv_i}	satellite clock offset for satellite i with respect to GPS time

The satellite positions and clock offsets are calculated from the navigation data transmitted by the satellites. Two different methods can be used to deny the full GPS accuracy: 1) manipulation of the navigation message orbit data, also referred to as the ϵ-process; and 2) manipulation of the satellite clock frequency, also referred to as the δ-process or clock dither. Manipulation of the navigation orbit data degrades the accuracy of the calculated satellite positions and results in slowly varying user position errors (periods on the order of hours). Note that the actual satellite orbits are not affected, only the parameters describing the satellite orbits are corrupted. Clock dither, on the other hand, involves the manipulation of the satellite clock itself. This results in fairly rapid errors on the pseudorange measurements with periods on the order of minutes. Because the actual satellite clock is manipulated, clock dither affects both the C/A code and the P code, as well as the integrated Doppler shift measurements.

Although position errors are specified in the GPS policy statement (see previous section), no information is provided with respect to the power spectral density of SA. Therefore, GPS receiver and system designers should anticipate a wide range of possibilities. The next section characterizes SA based on collected data.

III. Characterization of Selective Availability

The SPS provides information on the effects of SA in terms of position and time accuracies, which are summarized in Table 1. Also listed in Table 1 are the corresponding accuracies without SA, which were obtained from the U.S. Naval Observatory Bulletin Board.[3] The bulletin board also provided that frequency stability with SA would be on the order of 1 part in 10^{10}, whereas in the absence of SA, the frequency stability is on the order of 1 part in 10^{12}.

It should be noted that Table 1 provides position and timing accuracies only. No information is provided on the dynamics of the errors. The only way to

Table 1 Standard positioning service position and timing accuracies with and without selective availability

Parameter	With SA	Without SA
Horizontal position	100 m (95%) 300 m (99.99%)	20 m (95%)
Vertical position	140 m (95%)	30 m (95%)
Time	340 ns (95%)	40 ns

determine the dynamics of SA is through actual data collection. Figure 1 represents typical horizontal positioning accuracies of the GPS SPS with and without SA. Without SA, the horizontal position errors are fairly constant over periods on the order of tens of minutes. Position changes without SA are mostly caused by slowly changing propagation delays and satellite clock and ephemeris errors. With SA, the horizontal position "wanders around" within a circle with a radius of approximately 100 m. Successive position errors become uncorrelated after a period of approximately 2–5 min. Similar performance characteristics can be found for vertical positioning and for time transfer.

Although the characterization of SA in the position domain provides helpful information, several shortcomings of just a position domain characterization are quickly recognized.

1) SA is generated in each satellite and seems to be uncorrelated between satellites,[4] which means that the effect on the position accuracy depends on the satellite geometry.

2) The SPS policy assumes that at least 21 satellites are operational; therefore, additional satellite failures could produce larger than normal position errors.

3) No information is provided on velocity and acceleration errors.

4) Differential positioning accuracies are difficult to analyze with only position domain information. These depend on the acceleration of SA, processing time, and the latency of the data link.

5) No information is provided on the power spectral density of SA, which makes it difficult to simulate SA for receiver design and performance analysis purposes.

Because of these shortcomings, a complete analysis of SA-induced errors must start in the range measurement domain. Next, the range domain errors can be converted into positioning and timing errors depending on the application. Figure 2 shows representative, measured SA errors in the range domain. The SA error consists of the sum of a bias component (epsilon error) and a rapidly varying

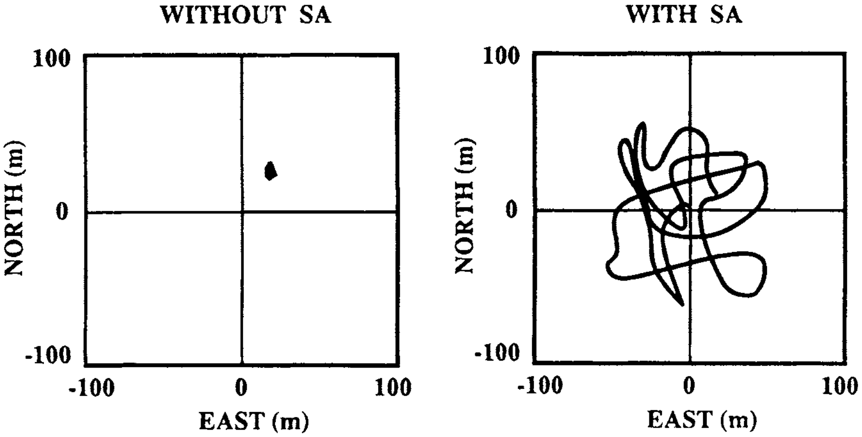


Fig. 1 Horizontal positioning errors with and without selective availability for data collected during a 1-h period.

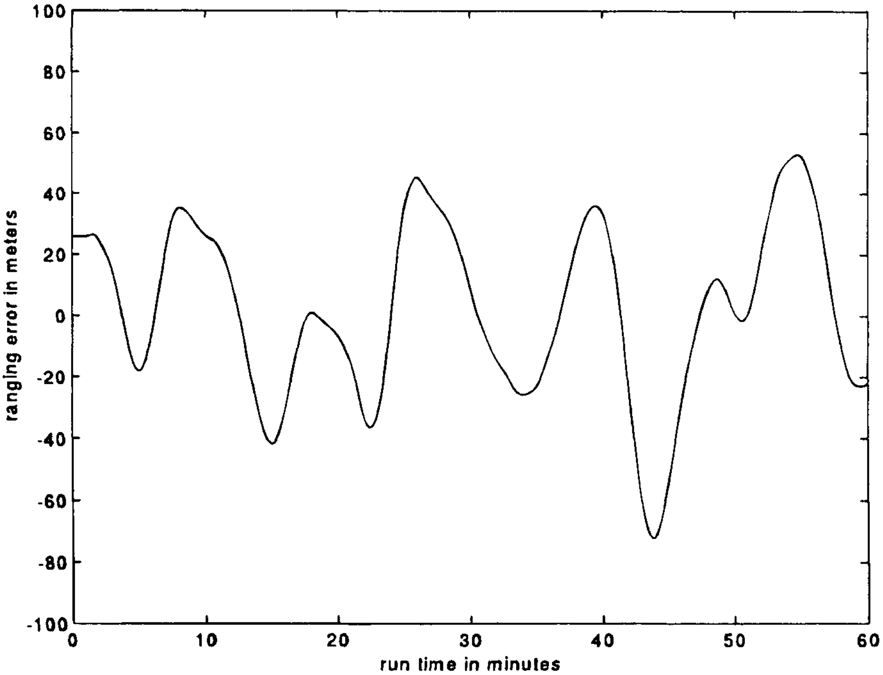


Fig. 2 Representative, measured range selective availability errors.

component (clock dither). The period of the oscillations is on the order of 2–5 min, while the standard deviation is approximately 23 m.

A first-order analysis of SA-induced ranging errors would start with the determination of the one-sigma range, velocity, and acceleration errors. Table 2 provides typical numbers resulting from such an analysis for two cases.[5] Case 1 refers to data collected from PRN 14 on day 124 of 1989, while case 2 represents data collected from both PRN 2 and 14 on day 259 of 1989. The user range accuracies (URA) were set to 32 and 64 m, respectively. The data without SA were collected from PRN 6 (a Block I satellite), while its URA was set to 2–4 m. Kremer et al.[5] gives detailed description of the data-processing techniques used. For both cases, the correlation time of the SA errors was on the order of 180 s. The range biases could either be caused by slow clock dither or by epsilon error.

Table 2 Example of measured line-of-sight range, velocity, and acceleration with and without selective availability

Parameter	One-sigma range	One-sigma velocity	One-sigma acceleration	Range bias
Without SA	1–1.5 m	0.0055 m/s	0.4 mm/s^2	0 m
With SA (case 1)	29 m	0.12 m/s	2 mm/s^2	74 m
With SA (case 2)	38–57 m	0.21 m/s	3.5 mm/s^2	−61–38 m

Further characterization of SA usually focuses on the clock dither. Epsilon errors are effectively simulated by adding random biases to the satellite orbit data. It should be noted, however, that SA could consist of either clock dither or epsilon errors, or both. RTCA, Inc. (formerly known as the Radio Technical Commission for Aeronautics), for instance, simulates epsilon error by adding random biases chosen from a Gaussian distribution with zero mean and a standard deviation of 23 m to the satellite range measurements. These biases are held constant during each GPS receiver test. Methods for identifying the epsilon and dither components of SA can be found in Ref. 6. Epsilon error may be found by comparing the orbits obtained from the broadcast ephemeris and from the postfit precise ephemeris. Dither is obtained by processing the corrections generated in a differential ground reference station.

The next step in the characterization of SA would be to determine the actual power spectral density (PSD) of the signal. This involves the postulation of a model for SA, which then can be used to generate statistically equivalent SA. The first SA model was derived by Matchett in 1985.[7] This model was not derived from actual data, but was deduced from a sample probability distribution curve. The GPS Joint Program Office (JPO) generated SA samples and then computed the distribution curve from these samples. A second-order Gauss–Markov process was postulated, and the coefficients were adjusted until its distribution curve matched the one provided by the JPO. A second-order Gauss–Markov process has also been adopted by Special Committee 159 of RTCA for the purpose of simulating SA errors.[8] The first models obtained from actual SA data were time series models derived by Braasch in 1989 using system identification theory.[9] The resulting models were autoregressive moving average (ARMA) models. Later, Chou implemented a second-order Gauss–Markov process based on measured SA data.[10] At the same time, Chou also provided the formulation for a recursive autoregressive model.[11] Other models of interest were published by Lear et al., who presented several time series and analytical models, also based on measured SA data.[12] Four of the above models are discussed in the following sections. These are the second-order Gauss–Markov model, an autoregressive (AR) model, an analytic model, and a recursive AR model (lattice filter). It should be noted that all of these models simulate clock dither only.

A. Second-Order Gauss–Markov Model

The second-order Gauss–Markov model implementation follows the description provided in Ref. 8. The continuous time model of a second-order Gauss–Markov process is given by the following:

$$\ddot{x}_p + 2\beta\omega_0\dot{x}_p + \omega_0^2 x_p = c \times w \tag{2}$$

where ω_0 = natural frequency; β = damping factor < 1; w = white Gaussian noise; and power spectral density = 1.

The continuous time model given by Eq. (2) can be expressed in the form of a state-space equation:

$$\begin{bmatrix} \dot{x}_p \\ \dot{x}_v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\beta\omega_0 \end{bmatrix} \begin{bmatrix} x_p \\ x_v \end{bmatrix} + \begin{bmatrix} 0 \\ c \end{bmatrix} w \tag{3}$$

where $\dot{x}_v = \dot{x}_p$. The corresponding discrete-time state-space equation is then given by the following:

$$\begin{bmatrix} x_p \\ x_v \end{bmatrix}_{i+1} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} x_p \\ x_v \end{bmatrix}_i + \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}_i \tag{4}$$

where w_1 and w_2 are white Gaussian noise processes with zero mean and unit variance. The elements of the discrete-time state transition matrix are as follows:

$$\begin{aligned} \phi_{11} &= e^{-\beta\omega_0\Delta T}[\cos(\omega_1\Delta T) + \beta(\omega_0/\omega_1)\sin(\omega_1\Delta T)] \\ \phi_{12} &= (1/\omega_1)e^{-\beta\omega_0\Delta T}[\sin(\omega_1\Delta T)] \\ \phi_{21} &= -\omega_0^2\phi_{12} \\ \phi_{22} &= e^{-\beta\omega_0\Delta T}[\cos(\omega_1\Delta T) - \beta(\omega_0/\omega_1)\sin(\omega_1\Delta T)] \end{aligned} \tag{5}$$

where ΔT is in units of seconds.

The error covariance matrix of the white noise is as follows:

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} \begin{bmatrix} u_{11} & 0 \\ u_{12} & u_{22} \end{bmatrix} \tag{6}$$

in such a way that

$$u_{11} = \sqrt{Q_{11} - Q_{12}^2/Q_{22}} \qquad u_{12} = Q_{12}/\sqrt{Q_{22}} \qquad u_{22} = \sqrt{Q_{22}} \tag{7}$$

The elements of the covariance matrix Q are as follows:

$$\begin{aligned} Q_{11} &= \frac{c^2}{4\beta\omega_0^3}\left[1 - \frac{\omega_0^2}{\omega_1^2} e^{-2\beta\omega_0\Delta T}\left(1 - \beta^2\cos(2\omega_1\Delta T) + \beta\frac{\omega_1}{\omega_0}\sin(2\omega_1\Delta T)\right)\right] \\ Q_{12} = Q_{21} &= \frac{c^2}{4\omega_1^2}[e^{-2\beta\omega_0\Delta T}(1 - \cos(2\omega_1\Delta T))] \\ Q_{22} &= \frac{c^2}{4\beta\omega_0}\left[1 - \frac{\omega_0^2}{\omega_1^2} e^{-2\beta\omega_0\Delta T}\left(1 - \beta^2\cos(2\omega_1\Delta T) - \beta\frac{\omega_1}{\omega_0}\sin(2\omega_1\Delta T)\right)\right] \end{aligned} \tag{8}$$

where

$$\omega_0 = \frac{\sigma_v}{\sigma_x} \qquad c^2 = 4\beta\frac{\sigma_v^3}{\sigma_x} \qquad \omega_1 = \frac{\sigma_v}{\sigma_x}\sqrt{1 - \beta^2} \tag{9}$$

$$\text{time constant: } \tau_d = \frac{1}{\beta}\frac{\sigma_x}{\sigma_v} = \frac{1}{\beta}\frac{1}{\omega_0}$$

$$\text{range sigma: } \sigma_x = \sqrt{\frac{c^2}{4\beta\omega_0^3}} \tag{10}$$

$$\text{velocity sigma: } \sigma_v = \sqrt{\frac{c^2}{4\beta\omega_0}}$$

The power spectral density function specified by RTCA for the above model is given by the following:

$$S(\omega) = c^2/(\omega^4 + \omega_0^4)\ \text{m}^2/(\text{rad/s}) \tag{11}$$

where ω is the frequency in rad/s.

The power spectral density is factored into right-half plane and left-half plane components:

$$S(\omega) = c/[\omega^2 - (2/\sqrt{2})(\omega_0\omega) + \omega_0^2]\ c/[\omega^2 + (2/\sqrt{2})\ (\omega_0\omega) + \omega_0^2] \tag{12}$$

Next, a second-order Gauss–Markov process is created by passing zero mean, unit variance, white Gaussian noise through the filter given by the left-half plane component of the power spectral density. The corresponding realization of the second-order Gauss–Markov process in the time domain is given by Eq. (2), see Ref. 13. From Eq. (2), it is found that the damping factor is given by the following:

$$\beta = 1/\sqrt{2} \tag{13}$$

For simulation purposes, RTCA has proposed the following parameters:

$$\begin{aligned} \omega_0 &= 0.012\ \text{rad/s} \\ c^2 &= 0.002585\ \text{m}^2 \end{aligned} \tag{14}$$

The model then results in the following:

$$\begin{aligned} \sigma_p &= \sqrt{\frac{0.002585}{4\frac{1}{\sqrt{2}}0.012^3}} = 23\ \text{m} \\ \sigma_v &= \sqrt{\frac{0.002585}{4\frac{1}{\sqrt{2}}0.012}} = 0.28\ \text{m/s} \\ \tau &= \sqrt{2}\,\frac{23}{0.28} = 118\ \text{s} \end{aligned} \tag{15}$$

The process outlined in the preceding equations is to be generated for each satellite. The error x_p is the second-order Gauss–Markov process to be added to the "perfect" pseudorange. The error x_v is the first-order Gauss–Markov process to be added to the instantaneous "perfect" pseudorange rate. Initialization of x_p is Gaussian with σ_p, initialization of x_v is Gaussian with σ_v.

For an update period of 1 s, the matrices in Eq. (4) become the following:

$$\Phi = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} = \begin{bmatrix} 0.9999284 & 0.9915387 \\ -0.0001428 & 0.9831014 \end{bmatrix}$$
$$U = \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} = \begin{bmatrix} 0.0146771 & 0.0252060 \\ 0 & 0.0504133 \end{bmatrix} \tag{16}$$

Figure 3 shows typical results for the second-order Gauss–Markov model. The range SA is slightly noisier than the actual range SA (see Fig. 2). The range rate SA is adequate for simulation purposes, but it is much noisier than actual range rate SA.

B. Autoregressive Model

This section follows the material presented in Ref. 6. In general, time series models are based on the assumption that the data of interest can be modeled as the output of a linear system (pole-zero filter) driven by Gaussian white noise. Conceptually, the derivation of a time series SA model can be thought of as a two-step process. The first step is to send the SA data through a filter and adjust the poles and zeros, or equivalently, the filter coefficients, in such a way that the output is Gaussian white noise with minimum variance (the output is referred

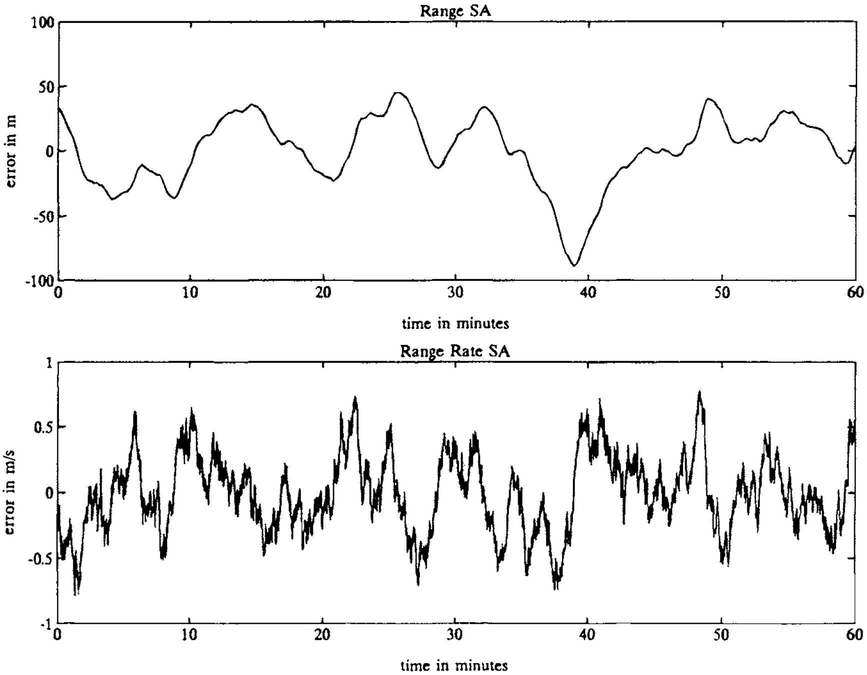


Fig. 3 Selective availability range and range rate errors generated by the second-order Gauss–Markov model.

to as residuals). The second step is to compute the inverse of the filter determined in the first step. Model identification is now complete.[6] Statistically equivalent SA data can then be generated by driving the inverse filter with Gaussian noise, whose variance is equivalent to that of the residuals in the first step. This process is illustrated in Fig. 4. Kelly provides and excellent overview of time series model identification and its application to the problem of microwave landing system (MLS) signal modeling.[14]

Three decisions are inherent in the procedure described in the preceding paragraph. The first is the choice of model (filter) type. Three are possible: 1) a pole-zero filter giving rise to an ARMA model; 2) an all-pole filter yielding an AR model; and 3) an all-zero filter yielding a moving average (MA) model. The second decision is the choice of model order. For example, if an AR model is chosen, how many poles will be used? The third decision is related to the first two and involves determining if a given residual sequence is white.

Because the primary goal is to derive an accurate SA-only model, an AR model type is chosen. This is because ARMA and MA models tend to be noisy. In fact, Braasch[9] concluded that an ARMA model was the best type for the combination of SA and receiver noise. An AR model of order p [referred to as an AR(p)] is defined as follows[15]:

$$x(n) = -\sum_{k=1}^{p} a(k)\, x(n-k) + e(n) \qquad (17)$$

where x is the model output; n is the time index; $a(k)$ is the kth filter coefficient;

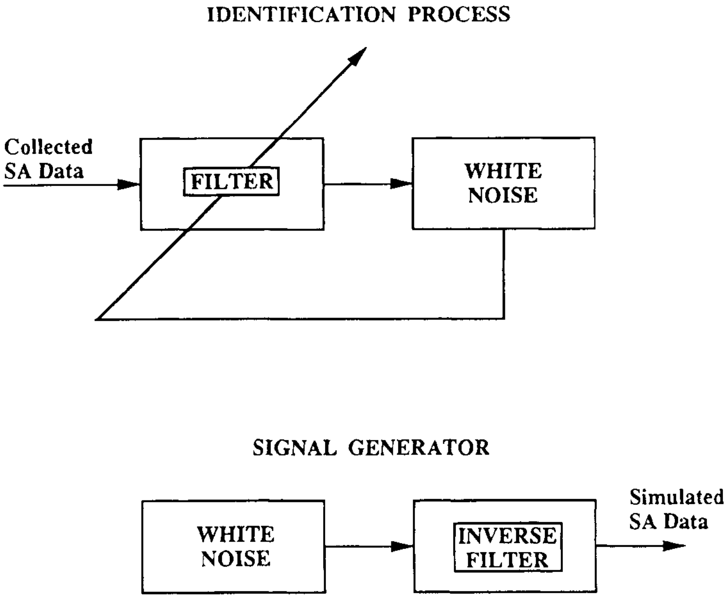


Fig. 4 Characterization of selective availability using system identification theory.

and e is the input Gaussian white noise. Note that SA models derived from data will operate at the same rate as the data collection rate.

Once having made the decision to use an AR model type, the rest of the process involves finding the optimum model order and coefficients (pole locations). For a given model order, many methods exist for optimizing the coefficients.[15–17] The one chosen is the modified covariance or forward–backward method. The second name stems from the fact that the optimization criterion is the minimization of forward and backward prediction errors. As shown later, this method performs quite well with SA data.

Several methods exist for model order selection. The majority of these have been developed for extremely short data records. The main issue is that we want to derive a model for the underlying statistical process that gave rise to the data. When model orders are selected that are too high (i.e., approaching the number of data points in the sample), the result is a "fit" of the sample data record rather than the underlying statistical process. The model order selection method used in this study is known as the Principle of Parsimony—the simplest acceptable model is the one chosen. An acceptable model is the inverse of the filter that outputs white noise when driven with SA. Note that if the model order is too low, the residuals will not be white, although the coefficients have been optimized.

The model identification, therefore, proceeds as follows. For a given sample of SA data, the coefficient is optimized for a first-order filter, and the residuals

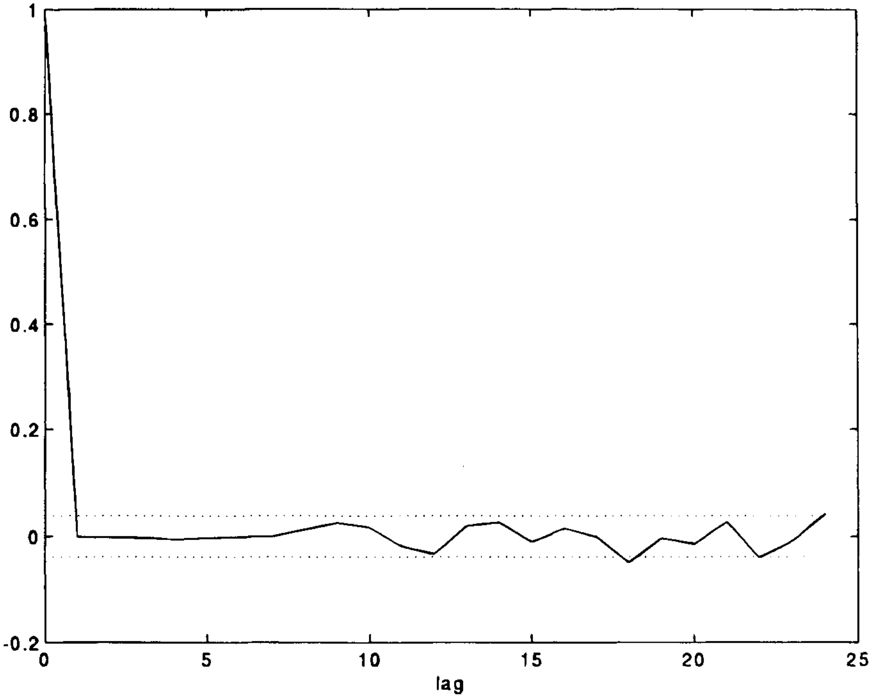


Fig. 5 Autocorrelation function of residuals for satellite 28.

are examined. If they are not white, then the coefficients for a second-order filter are optimized, and the residuals are examined again. The process is repeated until the model order and optimum coefficients are found for which the residuals are white. This process was performed for a variety of SA data sets. Depending upon the data set, models of either 9 or 11 coefficients were derived.

The method for determining whiteness involves examination of the autocorrelation function. An example is given in Fig. 5 where the autocorrelation function is plotted for the residuals from the SA data of satellite 28. Ideally, the autocorrelation function of white noise has a spike at lag zero and is zero everywhere else. However, this can be obtained only for infinite length sequences. As a result some minor "sidelobes" will occur at lags other than zero for white noise sequences that are finite. The dotted lines in the figure represent the 99% confidence intervals for the sidelobes. As can be seen in the plot, the sidelobes lie inside the confidence intervals for the most part, and thus the model is acceptable.

Further validation of the model can be performed by generating some waveforms and comparing the PSDs of the generated and collected data. An example is shown in Figs. 6 and 7. Figure 6 shows the waveform generated by the SA model derived from the SV 28 data. Note that if we compare the waveform to that of the collected data (Fig. 2), they are not the same; however, they are statistically equivalent. That is, the periods and amplitudes of the generated data are the same as for the collected data. This is better illustrated in Fig. 7 where the PSDs of the two waveforms are plotted. The lower line represents the collected

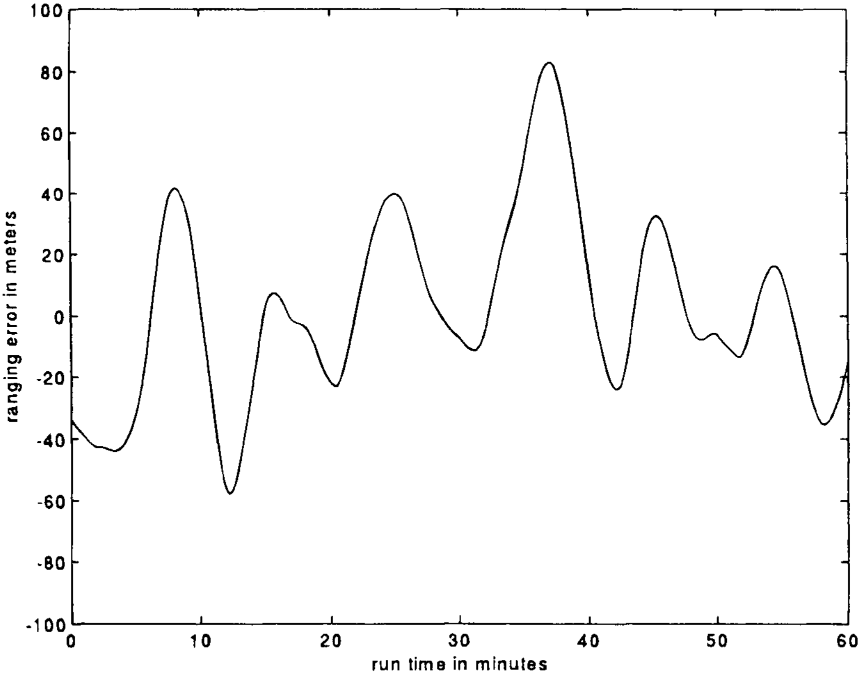


Fig. 6 Example selective availability model output for satellite 28.

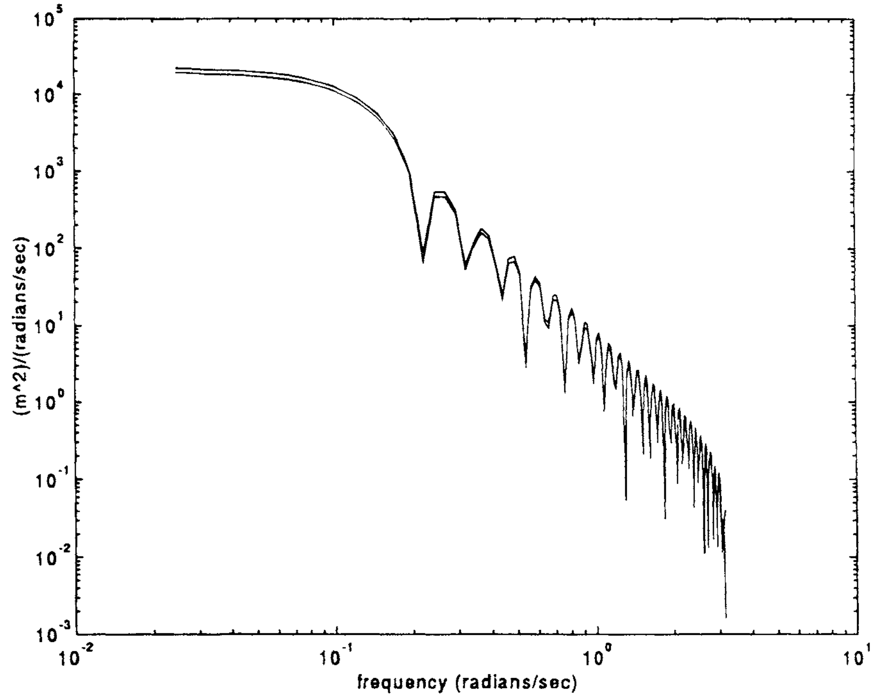


Fig. 7 Power spectral density functions of modeled and measured selective availability for satellite 28.

data, and the upper line represents the generated waveform. Power spectral density comparisons were performed on all of the models derived from the data. In each case, the result was similar to that shown here.

A final step in model validation concerns the power in the residuals. Recall that in step one of the model derivation process, the goal was to find a filter that output white noise (residuals) with minimum variance when driven with SA. The need for minimum variance is important from both a theoretical and practical viewpoint. Theoretically, having residuals with minimum variance means that the filter has been optimized and embodies the structure (i.e., correlation or information) of the SA. Kelly[14] refers to this as the filter "explaining" the data. However, from a practical viewpoint, minimum variance is also required. This is particularly true when trying to model random, yet smooth, waveforms such as SA.

Figures 8 and 9 illustrate the success of the AR model type in this respect. The residuals plotted in Fig. 8 have a standard deviation of 4.12 mm (4.12×10^{-3} m). Because this represents the amplitude of the noise driving the model [see Eq. (17)], it follows that any noise-like behavior in the generated SA waveforms will be negligible. This is verified in Fig. 9, which shows the smooth waveform of the generated SA over a short time interval.

A typical set of AR coefficients and the variance of the white noise input for

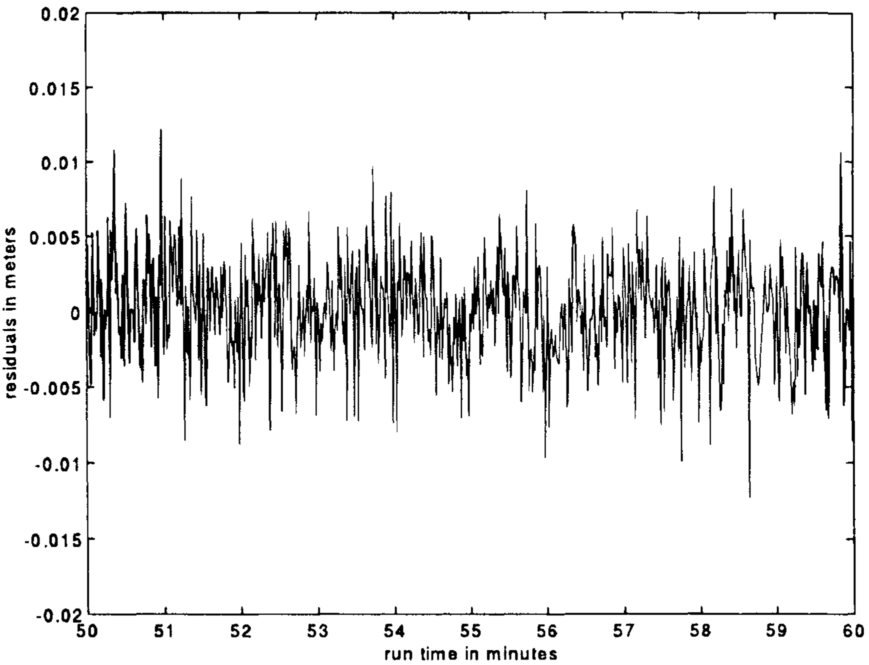


Fig. 8 Residuals for satellite 28.

data collected at a one-second update rate are given below:

$$a(1) = -1.36192741558063$$
$$a(2) = -0.15866710938728$$
$$a(3) = +0.13545921610672$$
$$a(4) = +0.21501267664869$$
$$a(5) = +0.30061078095966$$
$$a(6) = -0.12390183286070$$
$$a(7) = +0.10063573000351$$
$$a(8) = +0.02694677520401$$
$$a(9) = -0.12898590228866$$
$$a(10) = +0.05083106570666$$
$$a(11) = -0.05600186282898$$

$$\sigma_e^2 = 1.6993 \times 10^{-5}\ (\mathrm{m}^2)$$

where σ_e^2 is the variance of the Gaussian white noise input. These particular coefficients were derived from data collected from SV 28 during the first week of December 1992. Data from other satellites were found to produce similar results. The seemingly excessive number of significant figures are required to ensure filter stability.

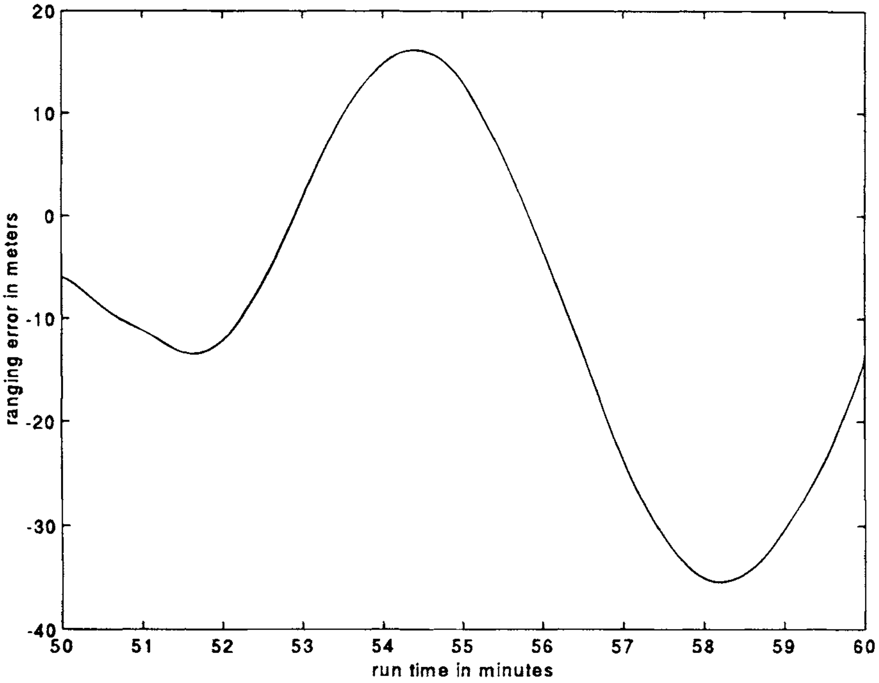


Fig. 9 Modeled selective availability for satellite 28—expanded scale.

C. Analytic Model

Lear et al. present several time series and analytical models in their 1992 paper.[12] Only one of the analytical models is discussed here because of its uniqueness. The term *analytic model* refers to "a piecewise smooth differentiable stochastic process with randomness appearing through the coefficients or model parameters".[12] These parameters are kept constant for randomly selected time intervals. The Rater analytic SA model was developed by observing data from satellite 19 on day 71 of 1992. Using low-noise integrated Doppler shift measurements and ionospheric corrections, the observed raw range was differenced twice. Data differenced once provided smooth pseudorange rate data, while the data differenced twice provided noisy pseudorange acceleration data. Rater observed that the second differences resemble connected ramps of varying slopes. This would imply that the third derivative (jerk) would consist of random steps. The third difference was too noisy, however, to show this behavior.

The Rater analytic SA model uses two randomly selected time intervals from uniform distributions. A constant jerk value is calculated for each time interval based on the length of the interval. The jerk is then integrated three times to obtain the range domain errors. Next, two of these waveforms (independently generated) are added together to obtain the range SA. In detail, the Rater model proceeds as follows:

1) Generate $x_1(t)$: Select the time duration T_1 of x_1 from a uniform distribution (100,700 s). Next, select the time constant τ_1 from a uniform distribution (0,

$T_1/2$ s) where τ is the period of time during which the jerk level is zero. This period is preceded by a jerk level of $+A_1$ and followed by a jerk level of $-A_1$. The jerk level A_1 is calculated from the following:

$$A_1 = -\ddot{x}_1(0)(2T_1 + \tau_1)/T_1(T_1 + \tau_1) \tag{18}$$

where $x_1(0) = \dot{x}_1(0) = 0$, and $\ddot{x}(0) = 4$ mm/s^2

The following jerk profile is generated:

Time interval	Jerk level
$0 < t \leq (T_1 - \tau_1)/2$	$+A_1$
$(T_1 - \tau_1)/2 < t \leq (T_1 + \tau_1)/2$	0
$(T_1 + \tau_1)/2 < t \leq T_1$	$-A_1$

The jerk profile is integrated three times to obtain $x_1(t)$.

2) Generate $x_2(t)$: Same as the generation of $x_1(t)$, but use new random numbers for T_2 and T_2 and choose $\ddot{x}_2(0) = -\ddot{x}_1(0)$.

3) Add the two processes $x(t) = x_1(t) + x_2(t)$.

The heart of this model lies in the basic waveform obtained after integrating the jerk profile three times. The result is a Gaussian-like pulse that is entirely non-negative or nonpositive depending upon the sign of $\ddot{x}_1(0)$. Then $x_1(t)$ consists of a series of positive pulses with random lengths and amplitudes, and $s_2(t)$ is a series of negative pulses. Finally, the sum of $x_1(t)$ and $x_2(t)$ results in a SA-like waveform.

Figure 10 shows a typical output of the Rater analytic SA model. As noted by the author of the model, a direct analytic relationship between the choice of parameters and the sigma of the process does not exist.[12] Nevertheless, the model seems to generate representative SA data. In addition, the model generates smooth range rate data, which is not the case for the previous two models.

D. Recursive Autoregressive Model (Lattice Filter)

A recursive AR model allows for the estimation of the AR model coefficients in real time, rather than in (off-line) batch processing. In principle, this would allow the AR model to predict ahead, once the model has "learned" the coefficients. The degradation of the accuracy of the predictions depends on the correlation time of the data. Highly correlated data would allow the AR model to predict ahead for longer periods of time than in the case of data that exhibit a low serial correlation. A recursive least squares (LS) lattice filter was first used by Chou in 1990.[11] The results presented by Chou were very promising, but they have not been repeated with the same level of success for other sets of data (see also Ref. 18). However, the use of a recursive AR model shows significant accuracy improvements in the prediction of the waveform when significant SA accelerations are presents. In the absence of significant SA accelerations, the recursive AR model does not offer any improvement over a simple linear extrapolation.

The LS lattice algorithm lends itself best to the real-time estimation of the AR coefficients because of the inherent stability offered by the lattice structure.[19] Following Refs. 16 and 19, the steps needed to implement the LS lattice filter are given below. The reader is referred to Ref. 19 for a detailed description of LS lattice filters.

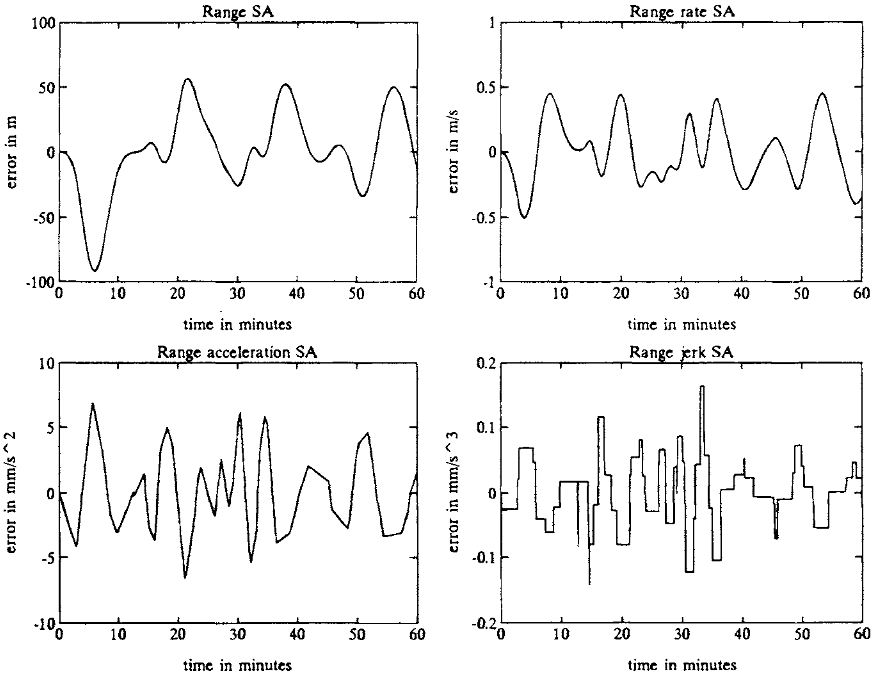


Fig. 10 Typical output of the Rater analytic selective availability model.

Initialization (δ is the steady-state squared prediction error; e.g., $\delta = 10^{-5}$):

$$e_m^b(0) = \Delta_m(0) = 0$$

$$\gamma_m(0) = 1 \qquad (0 \leq m \leq N)$$

$$\epsilon_m^f(0) = \epsilon_m^b(0) = \delta$$

For $n = 1$ to $n =$ final *do* (n is the current time index; x is the collected SA data)

$$e_0^b(n) = e_0^f(n) = x(n)$$

$$\epsilon_0^b(n) = \epsilon_0^f(n) = \epsilon_0^f(n - 1) + x^2(n)$$

$$\gamma_0(n) = 1$$

For $0 \leq m \leq N-1$ *do* (N is the model order)

The AR filter coefficients are calculated from

$$\Delta_{m+1}(n) = \Delta_{m+1}(n - 1) + [e_m^b(n - 1)e_m^f(n)/\gamma_m(n - 1)]$$

$$e_{m+1}^f(n) = e_m^f(n) - [\Delta_{m+1}(n)e_m^b(n - 1)/\epsilon_m^b(n - 1)]$$

$$e_{m+1}^b(n) = e_m^b(n - 1) - [\Delta_{m+1}(n)e_m^f(n)/\epsilon_m^f(n)]$$

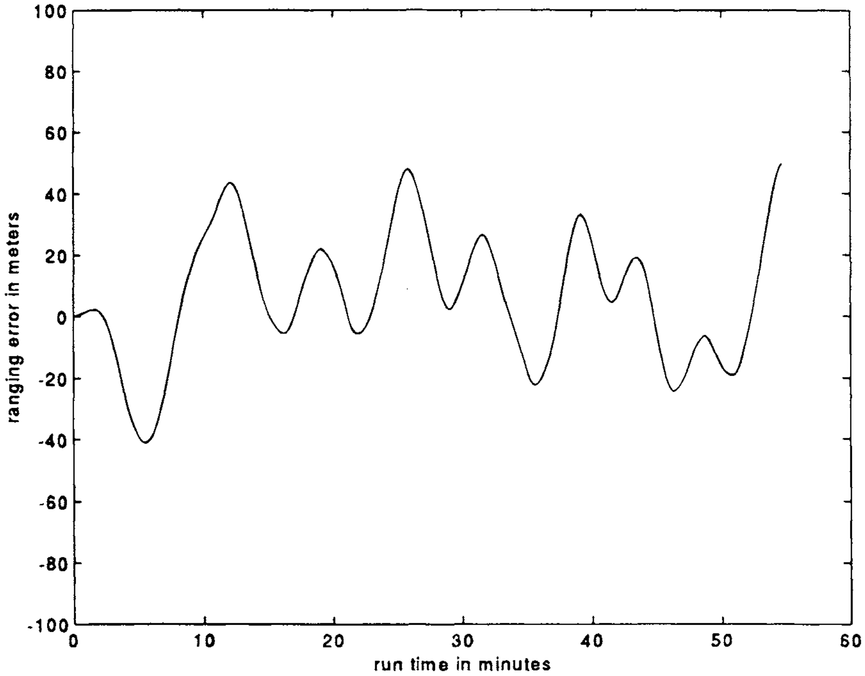


Fig. 11 Selective availability data used to "train" the least squares lattice filter.

```
        $\epsilon^f_{m+1}(n) = \epsilon^f_m(n) - [\Delta^2_{m+1}(n)/\epsilon^b_m(n - 1)]$
        $\epsilon^b_{m+1}(n) = \epsilon^b_m(n - 1) - [\Delta^2_{m+1}(n)/\epsilon^f_m(n)]$
        $\gamma_{m+1}(n - 1) = \gamma_m(n - 1) - \{[e^b_m(n - 1)]^2/\epsilon^b_m(n - 1)\}$
    The reflection coefficient is calculated from
        $k_{m+1}(n) = \Delta_{m+1}(n)/[\epsilon^b_m(n - 1)\ \epsilon^f_m(n)]^{1/2}$
        For $1 \leq i \leq m - 1$ do
            $\alpha_m(i) = \alpha_{m-1}(i) - k^b_{m+1}\alpha_{m-1}(m - i)$
        End ($i$ do loop)
            $\alpha_m(m) = k^b_{m+1}(n)$
    End ($m$ do loop)
    For $1 \leq i \leq N$ do
            $a_i(n) = \alpha_N\ (i)$
    End ($i$ do loop)
End ($n$ do loop)
```

The coefficients and parameters used in the LS lattice filter are summarized below:

$e^b_m(n)$ = the backward prediction error for order m at time equals n
$e^f_m(n)$ = the forward prediction error for order m at time equals n
$\Delta_m(n)$ = the partial correlation coefficient between the forward and backward prediction errors for order m at time equals n
$\gamma_m(n)$ = the angle parameter for order m at time equals n
$\epsilon^f_m(n)$ = the squared forward prediction error for order m at time equals n
$\epsilon^b_m(n)$ = the squared backward prediction error for order m at time equals n
δ = the steady-state squared prediction error
$x(n)$ = the collected SA data at time equals n
n = the current time index
N = AR model order
m = model order index (0 through $N-1$)
$k_m(n)$ = the reflection coefficient for order m at time equals n
$\alpha_m(i)$ = the ith prediction coefficient for order m
$a_i(n)$ = the ith AR filter coefficient at time equals n

As an example of the usefulness of the LS lattice filter, an 11th-order filter was "trained" with almost 1 h of simulated SA data generated with the model

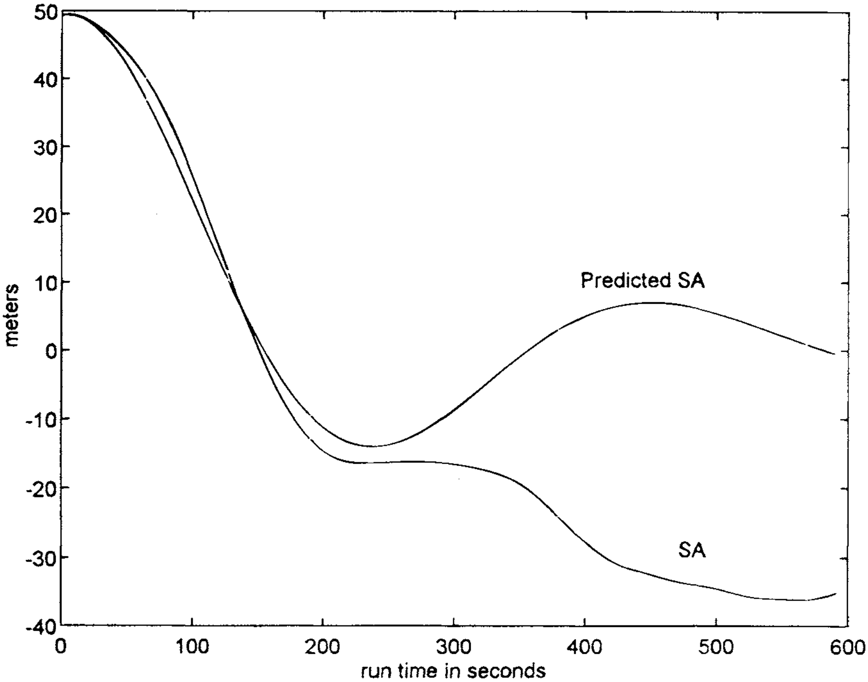


Fig. 12 Comparison of actual selective availability error and selective availability error predicted by the least squares lattice filter.

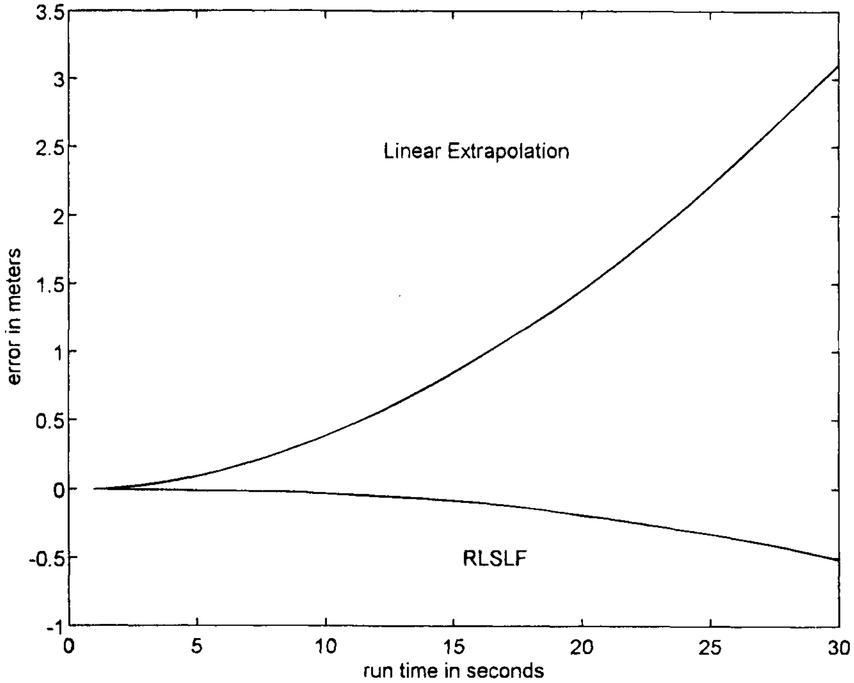


Fig. 13 Comparison of least square lattice filter prediction error and linear prediction error.

described in Sec. III. B., as shown in Fig. 11. At the end of the training period, the SA waveform was just starting to turn around from a positive slope to a negative slope. Figure 12 shows the predicted SA generated by the LS lattice filter, as well as the actual SA. During the first 3 min of the prediction interval, the predicted SA matches the actual SA to within 5 m. For periods of time longer than 3 min, the predicted values go to zero. This is consistent with the anticipated performance: the data can be predicted for a period of time on the order of the correlation time of the data. If a linear extrapolation is used instead of the LS lattice filter, the results are much worse, as illustrated in Fig. 13. After a period of 30 s, the linear extrapolation error is approximately 3 m, while the LS lattice filter prediction error is only −0.5 m. The opposite sign of the prediction errors is caused by the change in the slope of the SA data at the end of the training period. For this particular example, the LS lattice filter order was chosen to best match the SA data. Performance could be degraded if the filter is not adequately matched to the data. It should be noted that in the absence of significant SA acceleration, the performance of the linear extrapolator is statistically similar to the performance of the lattice filter.

E. Selective Availability Model Summary

Three different SA models are presented in this section. The highlights of each of the models are summarized as follows:

1. Second-Order Gauss–Markov Model

This model generates both range and range rate errors. It has been adopted by RTCA for GPS receiver-testing purposes. The main disadvantage of the model is that both the range and the range rate SA are noisier than that observed from actual satellite measurements.

2. Autoregressive Model

This model only generates range errors. Statistically, this is a very powerful model that is closely matched to measured SA data.

3. Analytic Model

This model generates both range and range rate errors. It also closely matches measured SA data. The range rate data represent the measured data better than those generated by the second-order Gauss–Markov model. The main disadvantage of this model is that there is not a direct relation between the choice of the model parameters and the output data. Noise levels must be set empirically.

Because the actual SA algorithms are not known, it is recommended, when required during GPS systems design and performance evaluations, to simulate the effect of SA using a variety of different models.

References

[1]Anon., "1992 Federal Radionavigation Plan," U.S. Department of Transportation and U.S. Department of Defense, Washington, DC, Rep. DOT-VNTSC-RSPA-92-2/DOD-4650.5, Jan. 1993.

[2]Georgiadou, Y., and Doucet, K. D., "The Issue of Selective Availability," *GPS World,* Sept.–Oct. 1990, pp. 53–56.

[3]United States Naval Observatory Electronic Bulletin Board, U.S. Coast Guard, Telephone (703) 866-3890, Aug. 29, 1990.

[4]Misra, P., Bayliss, E., LaFrey, R., and Pratt, M., "Integrated Use of GPS and GLONASS in Civil Aviation Navigation I: Coverage & Data Models," *Proceedings of the ION GPS-90,* Colorado Springs, CO, Institute of Navigation, Washington, DC, Sept. 1990, pp. 425–435.

[5]Kremer, G. T., Kalafus, R. M., Loomis, P. V. W., and Reynolds, J. C., "The Effect of Selective Availability on Differential GPS Corrections," *Navigation,* Vol. 37, No. 1, Spring 1990, pp. 39–52.

[6]Braasch, M. S., Fink, A., and Duffus, K., "Improved Modeling of GPS Selective Availability," *Proceedings of the ION National Technical Meeting,* San Francisco, CA, Institute of Navigation, Washington, DC, Jan. 20–22, 1993, pp. 121–130.

[7]Matchett, G., "Stochastic Simulation of GPS Selective Availability Errors," TM, FAA Contr. DTRS-57-83-C-00077, June 1985.

[8]Studenny, J., "Simulation of a Second-Order Gauss–Markov Process," RTCA Paper 148-93/SC159-424, March 17, 1993.

[9]Braasch, M. S., "A Signal Model for the NAVSTAR Global Positioning System," *Navigation,* Vol. 37, No. 4, Winter 1990–1991, pp. 363–377.

[10] Chou, H., "An Anti-SA Filter for Non-Differential GPS Users," *Proceedings of ION GPS-90,* Colorado Springs, CO, Institute of Navigation, Washington, DC, Sept. 19–21, 1990, pp. 535–542.

[11] Chou, H., "A Robust Filter for DGPS Users," *Proceedings of ION GPS-90,* Colorado Springs, CO, Institute of Navigation, Washington, DC, Sept. 19–21, 1990, pp. 607–612.

[12] Lear, W., Montez, M., Rater, L., and Zyla, L., "The Effect of Selective Availability on Orbit Space Vehicles Equipped with SPS GPS Receivers," *Proceedings of ION GPS-92,* Albuquerque, NM, Institute of Navigation, Washington, DC, Sept. 16–18, 1992, pp. 825–840.

[13] Åström, K., and Wittenmark, B., *Computer Controlled Systems—Theory and Design,* Prentice-Hall, Englewood Cliffs, NJ, 1984, pp. 146–151.

[14] Kelly, R. J., "MLS System Error Model Identification and Synthesis," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 28, No. 1, Jan. 1992, pp. 164–173.

[15] Marple, S. L., Jr., *Digital Spectral Analysis with Applications,* Prentice-Hall, Englewood Cliffs, NJ, 1988, pp. 172–260.

[16] Kay, S., *Modern Spectral Estimation: Theory and Application,* Prentice-Hall, Englewood Cliffs, NJ, 1987, pp. 106–270.

[17] Ljung, L., *System Identification—Theory for the User,* Prentice-Hall, Englewood Cliffs, NJ, 1987, pp. 169–207.

[18] Chou, H., "An Adaptive Correction Technique For Differential Global Positioning System," Ph.D. Dissertation, W. W. Hansen Experimental Physics Laboratory, Gravity Probe B Program, Stanford Univ., (SUDAAR 613), Stanford, CA, June 1991, pp. 59–82.

[19] Alexander, S. T., *Adaptive Signal Processing—Theory and Applications,* Springer-Verlag, New York, 1986, pp. 34–152.

Chapter 18

Introduction to Relativistic Effects on the Global Positioning System

N. Ashby*
University of Colorado, Boulder, Colorado 80309
and
J. J. Spilker Jr.†
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

A. Objectives

THE GPS is one of the first operational systems, outside of particle accelerators, that has important effects from relativity. The reasons for this are threefold. The GPS satellites have a large velocity, there is a non-negligible gravitational potential difference between that of the satellites and that of the users (usually at or near the Earth's surface), and there are significant Earth rotation effects. These effects of themselves are not necessarily important, clearly, there are other operational satellite systems. However, when coupled with the fact that GPS satellites carry precision atomic frequency standards, that pseudorange measurements can be made to accuracies in the nanosecond range, and that carrier-phase measurements are made to the centimeter level, relativistic effects can, indeed, be significant and must be taken into account (see also Chapter 3, this volume, and Ref. 1).

For a fixed user at sea level on the Earth's surface, there are three primary consequences of relativity effects (see Table 1):

1) There is a fixed frequency offset in the satellite's clock rate when observed from Earth. Most of the effect is purposely removed by slightly offsetting the satellite clocks in frequency prior to launch, the so-called "factory offset" of the clock.

2) The slight eccentricity of each satellite orbit causes an additional periodic clock error effect that varies with the satellite's position in its orbit plane.

Copyright © 1995 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Released to AIAA to publish in all forms.

*Professor of Physics.

†Ph.D., Chairman of the Board.

Table 1 List of key relativistic effects on GPS

- **Fixed user on the geoid**

  Constant frequency offset (factory offset)

  The net effect of relativity for a zero eccentricity GPS satellite is a combination of effects caused by the satellite's velocity and the Earth's gravitational field, including its quadrupole field. This effect produces a small fixed frequency offset of received signal frequency relative to the satellite transmitted frequency in addition to the classical Doppler shift. This effect is compensated by a prelaunch factory offset in the satellite clock.

  Sinusoidal delay perturbation caused by the eccentricity of the satellite orbit.

  Sagnac delay caused by the Earth's rotation during the time of transit of the satellite signal to the ground user.

- **Moving user or user above the geoid**

  In addition to the above effects, there are effects caused by the user velocity and height of the user above the geoid. Some of these effects can cancel or partially cancel in position estimation. These effects can be significant if the user is another satellite in orbit.

- **Secondary effects**

  There are a host of secondary effects discussed briefly in this chapter that are smaller than the accuracy level required by most users. These include:

  Tidal potential effects on clocks

  The user is on or near the rotating Earth, which in turn, is revolving about the sun with its gravitational field. However, both the satellite and user are in orbit about the sun at approximately the same position so much of this effect cancels.

  Nonspherical gravity potential

  The Earth's gravity potential is slightly nonspherical primarily because of the ellipsoidal shape of the Earth, which causes a currently nonmodeled quadrupole field effect on GPS satellite clocks.

  Shapiro delay

  The Shapiro delay is caused by variations in the apparent velocity of light because of the Earth's gravitational field. Approximate values of the Shapiro delay are calculated and shown to be small.

  Lense-Thirring effect–frame dragging

  This effect is caused by the rotation of the Earth's mass on its axis. This effect, a so-called frame-dragging effect, slightly modifies the solutions to the field equations and generates a slightly different metric with a term related to the Earth's angular momentum, but its effects are negligible for our purposes here.

3) There is also an effect (Sagnac delay) caused by the Earth's rotation during the time of transit of the satellite signal from satellite to ground.

Moving users on or near the Earth's surface or fixed users at an altitude above or below the geoid* may have to make additional corrections caused by their velocity and height above the ground. Satellite users in low-Earth orbit (LEO), geostationary-Earth-orbit (GEO), or other orbit altitudes have additional corrections that may be necessary depending upon the accuracy required. Clearly, the

* The geoid is an equipotential surface that differs slightly from the ellipsoidal model of the Earth by geoid undulations. The gravity potential W is the sum of both a gravitational potential Φ and a centrifugal potential $V = -{}^1\!/_2\,\omega^2(x^2 + y^2)$ and is also affected by nonuniformity in the Earth's density, and topography (see Sec. 3.1).

Table 2 Objectives of the chapter

- Introduce special and general relativistic effects. For the GPS, most of the significant general relativity effects can be shown from the principle of equivalence, and flat or nearly flat space models are adequate.
- Define the appropriate invariant equations, the metric tensors, and show their application to translational linear motion at constant velocity, rotating coordinates, spherically symmetric and cylindrically symmetric gravitational fields with and without coordinate system rotation.
- Compute the relativistic effects on GPS pseudorange and Doppler frequency measurements on both fixed and moving users, including users in orbit. Discuss effects on the calculation of user position, velocity, and time transfer estimates. Compute both primary and secondary effects.

impact of these relativistic effects can differ depending upon the objective and precision desired; e.g., for GPS and differential GPS (DGPS) position estimates, some common error relativistic effects cancel; whereas, for time transfer, they may not.

This chapter's objective is to describe for the reader the detailed effects of relativity for GPS users, and is summarized in Table 2. For an extensive treatment of relativity, the reader is referred to the texts in the References at the end of the chapter. In this chapter, the effects of the Earth's atmosphere are ignored. These effects have been discussed in Chapters 12 and 13, this volume, and are assumed to have been taken into account already.

B. Statement of the GPS Problem

As already discussed in detail in previous chapters, the GPS receiver makes two types of measurements on the received signal, pseudorange and carrier-phase/Doppler frequency measurements. In this chapter, we focus only on relativistic effects and assume that perfect atomic frequency standards are employed both for the satellite clock and the user clock. We ignore all other error sources covered in other chapters. Ideal atomic clocks moving with the user and satellite are sometimes referred to as *standard clocks.* For our purposes, it is assumed that the satellite transmits coded electromagnetic pulses at time instants as determined by this perfect satellite clock* and transmits these pulses at a precisely determined frequency f_o (see Fig. 1.). The satellite-generated coded pulse—really a sequence of pulses—carry embedded in the code the precise satellite clock time at the precise time of transmission of this coded pulse. The GPS receiver is also assumed to have a perfect atomic clock and measures the time of reception of the satellite-coded pulse waveform and decodes the satellite's clock time word. The receiver then simply subtracts the satellite clock time number (at the time of transmission) from the user clock time number at the time of reception. Corrections must be applied to account for the fact that the user undoubtedly

*The GPS clocks prior to launch are set to run at GPS time—an atomic standard time—with a factory offset for relativity. While in orbit, the satellite clocks are periodically corrected by the GPS Control Segment (CS) to match GPS time with both "paper" and, less frequently, physical clock corrections (see Chapters 4 and 10, this volume).

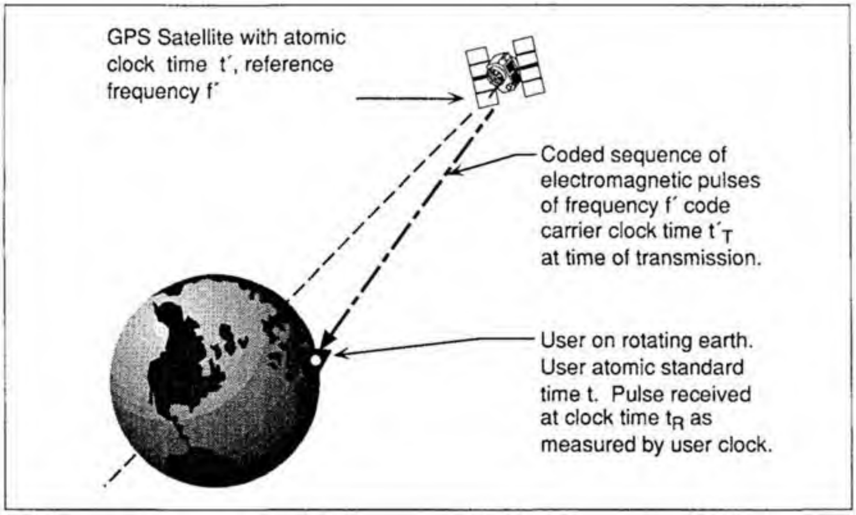

GPS Satellite with atomic clock time t′, reference frequency f′
Coded sequence of electromagnetic pulses of frequency f′ code carrier clock time t′$_T$ at time of transmission.
User on rotating earth. User atomic standard time t. Pulse received at clock time t$_R$ as measured by user clock.

Fig. 1 The GPS relativistic problem. The GPS satellite is assumed to carry a noise-free atomic frequency standard and produces a clock time t' carried on the satellite. The satellite transmits a coded sequence of electromagnetic pulses of frequency f'. The coded pulse sequence represents a number t'_T, which is the satellite standard clock time at the beginning of the coded pulse *transmission*. The GPS user receiver for this example also carries a perfect atomic frequency standard and produces time t. The receiver measures the time of *reception* t_R and then subtracts $t_R - t'_T$ to give pseudorange. The receiver moves during propagation of the signal. The receiver also measures the frequency of the received electromagnetic pulses and compares it with the frequency of its own coordinate clock to measure frequency shift.

has moved during the time of transit, because either the user is on the rotating Earth, in an aircraft slightly above the Earth's surface, or perhaps in another satellite in low-Earth orbit. Finally, the user also compares the satellite frequency and phase changes with its own clock to compute range rate and accumulated delta range (ADR), a carrier-phase measurement. In both instances, pseudorange and carrier frequency are affected by various relativistic effects. The problem is to account for motional Doppler shifts, gravitational frequency shifts, and propagation delays so the user can accurately determine his or her position and time.

The importance of relativistic effects for GPS, and especially for GPS time transfer, is heightened by the fact that today's atomic time standards carried in the GPS satellites and used on the ground are so precise. Today's clocks exhibit long-term Allan variance stability levels (see Chapter 4, this volume) on the order of 10^{-14}, and even better stability is possible for clocks on the ground.

Simultaneity is a crucial concept in the GPS. For users to determine position and time, atomic clocks in the satellites should be synchronized. For users to rendezvous successfully after traversing different paths with different speeds through different gravitational fields, their clocks must be synchronized, and their positions must be accurately determined. When coordinate reference frames are inertial, as in special relativity, the usual Einstein synchronization procedures may be employed.* However, typical users are moving over the rotating Earth and are in the Earth's gravitational field, giving rise to noninertial effects. Proper time elapsed on standard clocks depends upon the clocks' history, and the Einstein procedure is not sufficiently accurate. Instead, we use the concept of a *global coordinate time* wherein there is a hypothetical network of clocks synchronized in an underlying Earth–centered-inertial frame (ECI frame) and running at a rate consistent with atomic time standards on the geoid of the rotating Earth.[2] The definition of simultaneity employed is that of *coordinate simultaneity*,[3] namely, two events with space–time coordinates $\{t_1, x_1, y_1, z_1\}$ and $\{t_2, x_2, y_2, z_2\}$ are simultaneous in the given reference frame if $t_1 = t_2$.

In this chapter, standard time refers to *International Atomic Time* (TAI).†[4] Strictly speaking, GPS time is a coordinate time with rate determined by an independent composite set of clocks (see Chapter 10, this volume) which is indirectly coupled to TAI and Universal Coordinated Time (UTC) (with its leap seconds) through UTC–GPS corrections (see Chapter 4, this volume). Synchronization is achieved by methods discussed in this chapter.

II. Introduction to the Elementary Principles of Relativity

The mathematics of general relativity are extremely complex. However, for this treatment of the relativistic effects on GPS, only a small fraction of the theory is required, and even in this fraction, many simplifications can be made, because the gravitational fields (of the Earth) are relatively weak, and the velocities of the satellites and users are small compared to the velocity of light *c*. As a result, the curved spaces of general relativity are nearly flat.

Although much of the discussion of the relativistic effects of GPS involves either rotation of the Earth or satellites in orbit, and gravitational fields with some nonuniformity, we stress the admonition of Misner et al.[5] that "physics is simple when analyzed locally. Don't try to describe motion relative to far away objects." Physics is always and everywhere locally Lorentzian (has a local inertial frame). Thus, we begin the discussion by reviewing Euclidean geometry and inertial frames and then introduce the concepts of special relativity, the principle of equivalence and metric tensors. Metric tensors and their associated scalar intervals are the key tools that allow us to define the geometry of the space and interpret the consequences of the theory.

A. Euclidean Geometry and Newtonian Physics

In this section, we begin by considering families of inertial coordinate systems each moving at constant velocity (no acceleration) with respect to the others. To be called *inertial*, a coordinate system must satisfy the properties[6] of Table 3.

*Simultaneity can be defined self-consistently for a homogeneous, isotropic inertial frame; if light pulses emitted from two points, A and B, in the frame arrive at the midpoint of A and B at the same time, then the transmit events are simultaneous.

†TAI is a coordinate time scale, defined in a geocentric reference frame with the SI second as scale unit as realized on the rotating geoid. It can be extended to a fixed or moving point in the vicinity of the Earth with sufficient accuracy at the present state of the art by the application of first-order corrections of the General Theory of Relativity; i.e., corrections arising from differences in the gravitational potential and differences of speed, in addition to the rotation of the Earth.

Table 3 Required properties of an inertial coordinate system and related definitions

Event

An event marks an action that takes place at a point in space–time. Events are denoted by ϵ_1, ϵ_2, etc. for this chapter.

Observation by an inertial observer

An observer assigns to an event ϵ, the coordinates x, y, z of the location of the occurrence and the time t read by a clock in that coordinate frame located *exactly* at the position of that event for that frame.

World-line

A world-line is the path taken by a particle or light ray in four-dimensional space–time. The world-line is a line for which $\boldsymbol{x}$ is given as a function of t.

Inertial coordinate frame

The coordinate system must satisfy the three properties:

1) The distance between point P_1 at (x_1, y_1, z_1) and point P_2 at (x_2, y_2, z_2) is a constant independent of time.
2) The clocks that sit at every grid point in the coordinate frame are all synchronized, all run at the same rate, and provide a *coordinate time* for that coordinate frame, the same as proper time measured by synchronized standard clocks. *Proper time* is time measured on a single arbitrarily moving standard clock.
3) The geometry of space at any fixed instant of time is Euclidean, a flat space. Acceleration and gravitational fields violate property 2, except for a very localized region of space.

Much of our analysis focuses on these very localized regions of inertial space that apply to the GPS. Consider two coordinate systems S and S', where S' is moving at constant velocity v in the x direction relative to S, as shown in Fig. 2.

In standard Euclidean geometry, the *Galilean transformation* between coordinates for these Cartesian coordinate systems is as follows:

$$\begin{aligned} x' &= x - vt \\ y' &= y \\ z' &= z \\ t' &= t \end{aligned} \tag{1}$$

The very thought that t and t' could be different is not even considered in

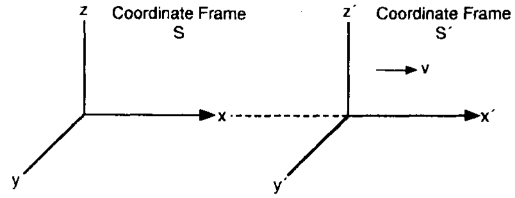


Fig. 2 Two (nonaccelerating) inertial coordinate frames S and S'. Coordinate frame S' is moving in such a way that $x' = x - vt$ where the origins of both frames are coincident at $t = t' = 0$.

Newtonian physics, and there is no constraint on velocity relative to the speed of light. In these Euclidean coordinate systems, the invariant differential interval between two events is the differential distance, given by the following

$$\begin{aligned} \mathrm{d}\sigma^2 &= \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 && \text{Cartesian coordinates} \\ &= \mathrm{d}r^2 + r^2\mathrm{d}\phi^2 + \mathrm{d}z^2 && \text{cylindrical coordinates} \\ &= \mathrm{d}r^2 + r^2(\mathrm{d}\theta^2 + \sin^2\phi\,\mathrm{d}\phi^2) && \text{spherical coordinates} \end{aligned} \qquad (2)$$

The *spatial metric* in Cartesian coordinates can be written as

$$\mathrm{d}\sigma^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \rho_{ik}\mathrm{d}x^i\mathrm{d}x^k = \delta_{ik}\mathrm{d}x^i\mathrm{d}x^k$$

where ρ_{ik} written in matrix form is the following diagonal matrix:

$$\|\rho_{ik}\| = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

See Table 4 for an explanation of the Einstein summation convention. Thus, a differential distance between any two points in Euclidean space is the same regardless of the chosen coordinate system; it is a scalar. This space is a perfectly flat space. The deficiency in this Galilean system is that it makes no allowance for the fundamental fact that the speed of light is a constant c in all inertial frames, nor that time measures could be different in different coordinate frames; i.e., t and t' can be different.

In special and general relativity where there are coordinate systems with different velocities, rotational effects, and gravitational effects, the three-dimensional spatial metric does not suffice. Instead, we deal with a four-dimensional space–time coordinate frame, and the *scalar interval* is of the form, $\mathrm{d}s^2 = g_{\alpha\beta}\mathrm{d}x^\alpha\mathrm{d}x^\beta$ where the individual $g_{\alpha\beta}$ terms can be functions of position and time; i.e., $g_{\alpha\beta}(x^\alpha)$. The set of elements $g_{\alpha\beta}$ together form a *covariant tensor* $g_{\alpha\beta}$ of rank 2. The tensor is symmetric; i.e., $g_{\alpha\beta} = g_{\beta\alpha}$. This *metric tensor* defines the geometry (differential geometry) of the space.* In general, the space–time coordinates are not flat. However, with the GPS and the users in relatively close proximity to the Earth, $r < 40{,}000$ km, the spatial coordinates are very nearly flat.

The fundamental basis for special relativity and its *Lorentz transformation* are contained in Einstein's postulates relating to the constancy of the speed of light in all inertial frames shown in Table 5.

B. Space–Time Coordinates and the Lorentz Transformation

Let us begin by considering four-dimensional space–time coordinates. This coordinate system may be envisioned as a standard Cartesian coordinate system filled with precision standard clocks at every point in the coordinate grid (see

*Any symmetric covariant tensor field of rank 2 defines a metric. A manifold with an applicable metric is termed a Riemanian manifold.[7] The concept of a differentiable manifold is abstract; however, it can be described roughly as an n-dimensional space that can be covered by one or more open neighborhoods, each of which has a coordinate system. Pairs of these coordinate systems are related to each other by differentiable coordinate transformations. A smooth surface in three-dimensional Euclidian space is a simple example of a differentiable manifold.[8]

Table 4. General expressions for distance in three-dimensional space and four-dimensional space–time

We introduce a set of three-dimensional spatial coordinates $\{x, y, z\} = \{x^1, x^2, x^3\}$ in frame S and a new set of coordinates $\{x'^1, x'^2, x'^3\}$ in frame S' where $x = x^1(x'^1, x'^2, x'^3)$, $y = x^2(x'^1, x'^2, x'^3)$, $z = x^3(x'^1, x'^2, x'^3)$. A coordinate differential $\mathrm{d}x = \mathrm{d}x^1$ is related to coordinate differentials $\mathrm{d}x'^i$ by the following:

$$\mathrm{d}x = \mathrm{d}x^1 = \frac{\partial x}{\partial x'^1}\mathrm{d}x'^1 + \frac{\partial x}{\partial x'^2}\mathrm{d}x'^2 + \frac{\partial x}{\partial x'^3}\mathrm{d}x'^3 = \sum_i \frac{\partial x}{\partial x'^i}\mathrm{d}x'^i = \frac{\partial x^1}{\partial x'^i}\mathrm{d}x'^i$$

Einstein's summation convention is used, where two repeated indices—one subscript and one superscript—automatically imply a summation over the range of the indices. Latin indices range from 1 to 3, while Greek indices range from 0 to 3. Thus, $a_i b^i$ is shorthand for $a_1 b^1 + a_2 b^2 + a_3 b^3$.

Similarly, the differentials for $\mathrm{d}y$ and $\mathrm{d}z$ may be written in the following compact form:

$$\mathrm{d}x^j = \sum_i \frac{\partial x^j}{\partial x'^i}\mathrm{d}x'^i = \frac{\partial x^j}{\partial x'^i}\mathrm{d}x'^i$$

Quantities that transform under a change of coordinates according to such a rule: $V^j = \partial x^j/\partial x'^i \; V'^i$ are termed *contravariant vectors.* (In three-dimensional Euclidean space, the distinction between contravariant and covariant vectors is not as significant as it is in general space-time.)

The invariant distance measure is as follows:

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 = \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j = \delta_{ij}\frac{\partial x^i}{\partial x'^m}\frac{\partial x^j}{\partial x'^n}\mathrm{d}x'^m\mathrm{d}x'^n$$

$$= g'_{mn}(x'^1, x'^2, x'^3)\mathrm{d}x'^m\mathrm{d}x'^n$$

where the *metric tensor* in the new coordinate system $\{x'^1, x'^2, x'^3\}$ is the following:

$$g'_{mn} = \delta_{ij}\frac{\partial x^i}{\partial x'^m}\frac{\partial x^j}{\partial x'^n} = \left[\frac{\partial x}{\partial x'^m}\frac{\partial x}{\partial x'^n} + \frac{\partial y}{\partial x'^m}\frac{\partial y}{\partial x'^n} + \frac{\partial z}{\partial x'^m}\frac{\partial z}{\partial x'^n}\right]$$

The three-dimensional distance measure, thus, simplifies to $\mathrm{d}s^2 = g'_{ij}\mathrm{d}x'^i\mathrm{d}x'^j$. The three-dimensional metric tensor g'_{ij} in an arbitrary coordinate system has nine components but only six are independent because $g'_{ij} = g'_{ji}$.

The metric tensor generalizes to four-dimensional spacetime using the coordinate x^0 for ct where t is the time for that coordinate system, and the spatial coordinates are $\{x^1, x^2, x^3\}$. Then a general *line element* in four dimensions is written

$$\mathrm{d}s^2 = g_{\alpha\beta}\mathrm{d}x^\sigma\mathrm{d}x^\beta$$

The particular dependence of $g_{\alpha\beta}$ on space–time is determined by the physical situation (see Table 10).

In special and general relativity, the increment of *proper time* $\mathrm{d}\tau$ elapsed on a standard clock that moves along the space–time path $\mathrm{d}x^\alpha$ is just $\mathrm{d}\tau = |\mathrm{d}s|/c$. Thus, if a clock does not move, $\mathrm{d}x^\alpha = 0$ for $\alpha = 1, 2, 3$, and then $\mathrm{d}\tau^2 = \mathrm{d}s^2/c^2 = g_{00}(\mathrm{d}x^0)^2/c^2$, and the proper time elapsed on the clock is $\sqrt{|g_{00}|}\,\mathrm{d}t$.

Table 5 Einstein's two postulates[9]

Postulate 1: The Principle of Relativity

"No inertial system is preferred. The equations expressing the laws of physics have the same form in all inertial systems."

Postulate 2: The Principle of the Constancy of the Speed of Light

"The speed of light is a universal constant independent of the state of motion of the source. Any light ray moves in the inertial system of coordinates with constant velocity, c, whether the ray is emitted by a stationary or by a moving source."[10]

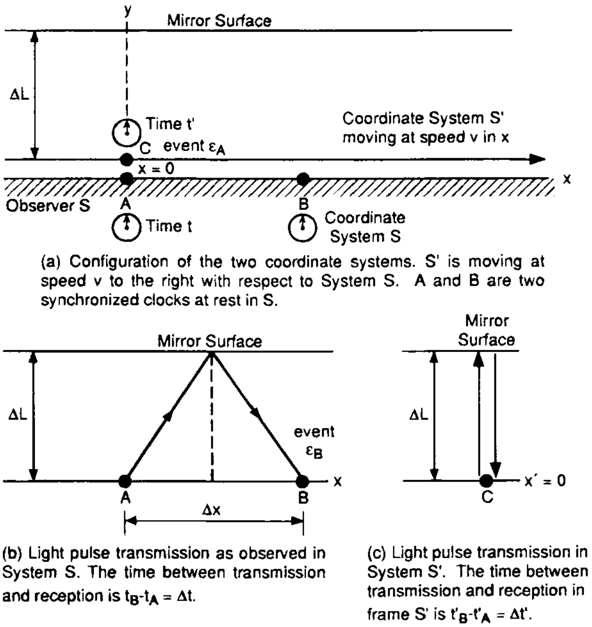
Of course, the speed of light c refers to the speed of light in a vacuum. Obviously, the speed of light decreases in a dielectic medium such as air or glass fiber.

Ref. 11). Second, define *events*; for example, a light pulse transmission as, say, ϵ_1, not only by the place of occurrence x, y, z, but also by the time t in that coordinate system at which the event took place. Thus, an event ϵ_1 is labeled in an inertial (nonaccelerating, gravity-free) coordinate frame S as $\epsilon_1(t_1, x_1, y_1, z_1)$ and in coordinate frame S' by $\epsilon_1(t_1', x_1', y_1', z_1')$ where the clock time is labeled by the time on the clock that is *exactly* at the same location in each coordinate system as the event. The two coordinate frames S and S' are generally in relative motion to one another. Each frame must have its own set of measuring rods and synchronized clocks at rest in that frame. An "observer" is a shorthand way of speaking of the whole ensemble of clocks and measuring rods associated with one coordinate frame, and an observation is the operation of labeling each event with the position coordinates and clock time at that event. It is critical that the clock time label be that associated with the exact position of that event, not a clock time at some distant point. An observation is *not* the act of looking at some clock distant from the event, and attempts to do so will often lead to error and sometimes leads to so-called paradoxes.

1. Time Dilation

Consider next two inertial coordinate frames, as shown in Fig. 3, where the origin of the S' frame moves with respect to reference frame S along the x-axis at velocity v, i.e., $dx/dt = v$. At time $t = 0 = t'$, the origins coincide. Consider an event ϵ_A wherein a light flash is emitted at the origin of both coordinate frames; i.e., ϵ_A $(t = 0, x = 0, y = 0, z = 0)$, ϵ_A $(t' = 0, x' = 0, y' = 0, z' = 0)$. Event B is the reception of the flash after reflection from a mirror or other reflector at position $y = \Delta L = y'$. (Symmetry considerations imply that the two observers agree on lengths oriented perpendicular to the direction of relative motion.)

Note that the speed of light c is exactly the same in both frames, from Postulate 2 in Table 5. Assume that the reflecting mirror moves with the coordinate frame S' (or equivalently, that the light beam is angled in the S frame so that it returns to the origin of S'). The light pulse in frame S' travels exactly $2\Delta L$ meters, and hence event ϵ_B in S' has coordinates $(t' = 2\Delta L/c, x' = 0, y' = 0, z' = 0)$. Thus, the coordinate changes between the two events in S' are $\Delta t' = 2\ \Delta L/c$, $\Delta x' = 0$, $\Delta y' = 0$, $\Delta z' = 0$.



(a) Configuration of the two coordinate systems. S' is moving at speed v to the right with respect to System S. A and B are two synchronized clocks at rest in S.

(b) Light pulse transmission as observed in System S. The time between transmission and reception is $t_B - t_A = \Delta t$.

(c) Light pulse transmission in System S'. The time between transmission and reception in frame S' is $t'_B - t'_A = \Delta t'$.

Fig. 3 Example of light pulse transmission in two coordinate frames *S* and *S'*. The light pulse is transmitted, event ϵ_A, at position *A* in frame *S* and at position *C* in frame *S'*, which is exactly coincident with *A* at the time of event ϵ_A. The pulse is received, event ϵ_B, at the same location *C* in frame *S'* and at a displaced point *B* in frame *S* where $\Delta x = v\Delta t$.

In coordinate frame *S*, on the other hand, the light pulse arrives, event ϵ_B, at the following time:

$$t = \Delta t = 2\sqrt{(\Delta L)^2 + (\Delta x/2)^2}/c \tag{3}$$

where $\Delta x = \text{v}\Delta t$, and $x = \Delta x$, $y = 0 = z$. Thus, in *S*, event ϵ_B has coordinates $t = \Delta t, x = v\Delta t$, $y = 0$, $z = 0$. The coordinate changes between the two events are, therefore, (Δt, $\Delta x = v\Delta t$, $\Delta y = 0$, $\Delta z = 0$). Because $\Delta x = v\Delta t$, we can solve Eq. (3) for Δt to obtain the following:

$$\Delta t = \frac{2\Delta L/c}{\sqrt{1 - (v/c)^2}} \tag{4}$$

The ratio of the two time intervals is as follows:

$$\frac{\Delta t}{\Delta t'} = \frac{1}{\sqrt{1 - (v/c)^2}} = \frac{1}{\sqrt{1 - \beta^2}} \stackrel{\Delta}{=} \gamma \geq 1 \tag{5}$$

where γ is the *time dilation* factor, and $\beta \stackrel{\Delta}{=} v/c$ is the normalized speed, with $v = |\mathbf{v}|$. Thus, the laboratory reference frame clock accumulates a larger clock count between the same two events than the clock in the moving coordinate system S', and hence, the moving clock seems to be running at a lower frequency clock rate than the laboratory clock. *Moving clocks run slow when compared with a sequence of synchronized clocks at rest.* Of course, if the experiment were reversed so that the light pulse returned to the origin of S, then $\Delta x = 0$, the clock at A in S is moving at velocity $-v$ with respect to S', and it would seem that the S frame clock is running slow.

For the two events ϵ_A, ϵ_B discussed here, the clock in S' moves uniformly from ϵ_A to ϵ_B, and in S', the events occur at the same position in space. Whenever it is possible to introduce a uniformly moving clock whose world line intersects two events, an inertial frame can be constructed with this clock as reference, in which the spatial separation between the events is zero. The space–time separation between the events is then *timelike*, and the proper time elapsed on the clock is a measure of this separation. We can then show that, in general, the space–time separation between events $\Delta s^2 = -(c\Delta t)^2 + \Delta x^2 + \Delta y^2 + \Delta z^2$ is a scalar quantity having the same value in all inertial frames. Indeed, in the present case in S' $-(c\Delta t')^2 + (\Delta x')^2 + (\Delta y')^2 + (\Delta z')^2 = -(c\Delta t')^2 = -4(\Delta L)^2$, while from Eq. (4) in S, $-(c\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 = (c\Delta t)^2(1 - v^2c^2) = [-4(\Delta L)^2/(1 - \beta^2)]\,(1 - \beta^2) = -4(\Delta L)^2$. Thus, because v no longer appears, Δs^2 is an invariant quantity independent of the coordinate frame.

In differential form* we write the following:

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \tag{6}$$

This invariant is termed the space–time interval for the *Lorentz space* (inertial space). Notice that if the clock is fixed in the reference frame; i.e., $\mathrm{d}x = \mathrm{d}y = \mathrm{d}z = 0$, then $\mathrm{d}t = \mathrm{d}\tau$, and $\mathrm{d}t$ is equal to proper time. Thus, a fixed (standard) clock in the coordinate system carries proper time (see Table 1). *Proper time* is the time carried by a clock that is transported along a world line in space–time through the events. Figure 4 shows a world-line description of the light pulse transmission. A *world line* is the path taken by a particle or light ray in space–time.

The expression for the invariant $\mathrm{d}s^2$ contains one form of the metric tensor and can be written for this inertial space as the following quadratic form†:

$$\mathrm{d}s^2 = \sum_{\alpha}^{3}\sum_{\beta}^{3} g_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta = g_{\alpha\beta}\,\mathrm{d}x^\alpha\,\mathrm{d}x^\beta \tag{7}$$

For inertial space (Minkowski space), the metric tensor in Cartesian coordinates can be written in matrix form as follows:

*Note that many texts use simplifying normalizing notations so that c does not appear, and some use $\mathrm{d}s^2 = -\mathrm{d}t^2 + (\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)/c^2$ and express the line element in time dimensions. In this chapter, $\mathrm{d}s$ has the dimensions of meters.

†The metric tensor $g_{\alpha\beta}$ is a tensor of rank 2 that carries out an operation $\boldsymbol{g}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{x}) = g_{\alpha\beta}\,\mathrm{d}x^\alpha\mathrm{d}x^\beta$ on any two vectors inserted and computes a scalar that has certain invariant properties under coordinate transformation. For more detail refer to Refs. 5, 8, 12, 13. We are using Einstein notation and automatically sum over any index that appears as both a subscript and superscript.

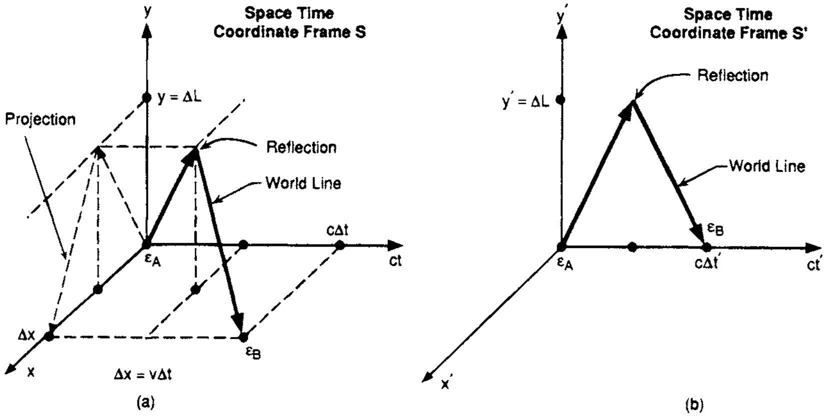


Fig. 4 Space–time coordinate frame world lines in coordinate frames S and S' showing events ϵ_A and ϵ_B. Coordinate frame S' is moving in the x direction with respect to coordinate frame S at speed v. A light pulse is transmitted, event ϵ_A, when the origins at time $t = 0 = t'$ at the origin of S and S'. The pulse is reflected at $y = \Delta L$ and returned to $y = 0$; $y' = 0$, event ϵ_B.

$$\|g_{\alpha\beta}\| = \|\eta_{\alpha\beta}\| \triangleq \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where $\eta_{\alpha\beta}$ is defined as the metric tensor for *Minkowski space* and is one form of the generalized Kronecker delta. This scalar invariant, or line element, can also be written in cylindrical and spherical coordinates as follows:

$$ds^2 = -c^2dt^2 + dr^2 + r^2d\phi^2 + dz^2$$
$$ds^2 = -c^2dt^2 + dr^2 + r^2(d\theta^2 + \sin^2 d\phi^2) \tag{8}$$

The line element for any path in space–time is invariant under coordinate transformations between inertial frames. Along the path of a photon, the line element or invariant distance measure, $ds^2 = 0$. Notice that because for photon travel $ds^2 = 0$, the line element yields the result $c^2dt^2 = dx^2 + dy^2 + dz^2$, which is an expression of the constancy of the speed of light in all inertial frames.*

2. *Geodesics*

A geodesic is the path in space–time that a body takes when it is free from nongravitational forces. In the flat space of special relativity, inertial space, the body moves with uniform speed along a straight line. In more general spaces,

*Note that in three dimensions x, y, t, this equation $c^2\,dt^2 = dx^2 + dy^2$ can be visualized as a cone in three dimensions, sometimes referred to as a light cone. Physical objects must travel inside the light cone, and light rays travel on the surface of the cone.

the geodesics* are extremal paths. For physical particles, a geodesic ordinarily corresponds to the shortest path. For light rays, the geodesics have zero space–time length and are termed null geodesics. In general, in a gravity field, the geodesics are curved, at least slightly, in space.

3. *Transformations Between Inertial Frames*

Consider an inertial system S with $\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. This space–time interval is not form-invariant under Galilean coordinate transformations, because if we apply Eqs. (1) to transform $\mathrm{d}s^2$, we obtain

$$\mathrm{d}s^2 = -(c\mathrm{d}t')^2\,(1 - \beta^2) + 2\beta\mathrm{d}x'c\mathrm{d}t' + \mathrm{d}x'^2 + \mathrm{d}y'^2 + \mathrm{d}z'^2,$$

which is not of the same form as Eq. (6). However, because it is a quadratic form in the primed coordinate increments, by performing additional linear transformations of x' and t' it is possible to reduce this expression to the appropriate form $-(c\mathrm{d}t')^2 + \mathrm{d}x'^2 + \mathrm{d}y'^2 + \mathrm{d}z'^2$.

Let us, instead, begin by assuming that x and ct are linear combinations of x' and ct'. For the event ϵ_B of Fig. 3, $x' = 0$, while $ct = \gamma ct'$, and $x = vt = \gamma\beta ct'$. This set of relationships is enough to determine two of the four needed coefficients. Then, for an arbitrary event we must have the following:

$$ct = \gamma ct' + Ax' \tag{9}$$

$$x = Bx' + \gamma\beta ct' \tag{10}$$

where A and B are constants that remain to be computed. Consider events at the origin of S, at $x = 0$. These must occur at $x' = -vt'$, so from Eq. (10) $0 = B(-vt') + \gamma\beta ct'$, which gives $B = \gamma$. Invariance of the speed of light implies that if $x' = ct'$, then $x = ct$. That is, the position of a light pulse transmitted from the origins at the moment they coincide must move with speed c in both coordinate systems. Then, substituting $x' = ct'$ in Eqs. (9) and (10) we must have $x = \gamma ct' + \gamma\beta ct' = ct = \gamma ct' + Act'$.

This equation can only be satisfied if $A = \gamma\beta$. The resulting transformations, the so-called Lorentz transformations, can be written as follows:

$$\begin{aligned} ct &= \gamma(ct' + \beta x') \\ x &= \gamma(x' + \beta ct') \end{aligned} \tag{11}$$

There are many other ways of deriving the Lorentz transformations. Another method is to assume the transformations between coordinates of S' and S are linear, then impose the requirement that $\mathrm{d}s^2$ must have the same form in the two frames. This yields Eqs. (11) after some algebra.

Compare Eq. (11) with the Galilean transformation where we would have $x = x' + vt'$ and $t = t'$. First-order corrections enter in the time transformation; second-order corrections enter through the factor γ.

*A geodesic is a curve for which the integral $\int \mathrm{d}s$ is an extremum; the curve has zero geodesic curvature.[11]

4. *Velocity Addition*

Consider a double transformation of velocities where there are three inertial coordinate systems: S_a, S_b, S_c, as shown in Fig. 5. The coordinate frame S_b moves at velocity v_b with respect to S_a, and S_c moves at velocity v_c with respect to S_b. The velocities would, of course, simply add in a Galilean transform with S_c moving with respect to S_a with velocity $v_b + v_c$. Let the three sets of coordinates be $\{ct_a, x_a, y_a, z_a\}$, $\{ct_b, x_b, y_b, z_b\}$, and $\{ct_c, x_c, y_c, z_c\}$.

Using the Lorentz transformation (11), we can write the following:

$$\begin{aligned} x_a &= \gamma_b(x_b + v_b t_b) = \gamma_b[\gamma_c x_c + \gamma_c v_c t_c] + \gamma_b v_b[\gamma_c t_c + v_c \gamma_c x_c] \\ &= \gamma_b\gamma_c[x_c + v_c t_c + v_b t_c + v_c v_b x_c] = \gamma_b\gamma_c[x_c(1 + v_b v_c) + (v_b + v_c)t_c] \\ &= \gamma_b\gamma_c(1 + v_b v_c)\left[x_c + \left(\frac{v_b + v_c}{1 + v_b v_c}\right)t_c\right] \end{aligned} \tag{12}$$

where $\gamma_b = 1/\sqrt{1 - \beta_b^2}$, and $\beta_b \triangleq v_b/c$, etc. Thus, the effective velocity of S_c with respect to S_a in the x direction is as follows:

$$v_{\text{eff}} = \frac{v_b + v_c}{1 + v_b v_c} \quad \text{and} \quad \gamma_{\text{eff}} \triangleq 1/\sqrt{1 - (v_{\text{eff}}/c)^2} = \gamma_b\gamma_c(1 + v_b v_c) \tag{13}$$

Thus, $v_{\text{eff}} \leq c$. This result is generalized in Table 6.

It is important to point out that this velocity addition applies equally when a coordinate system S_b, moving with respect to S_a, has a moving transmitter on it. That is, the transmitter is moving with respect to S_b at velocity v_c and is observed by S_a. We cannot simply use the Galilean velocity addition formulas except as an approximation for small v/c. Tables 7 and 8 summarize the four-velocity and four- momentum characteristics in inertial space. Table 9 summarizes the definition and symbols used herein.

5. *Light Pulse Time Delay*

Now we consider how different observers, in relative motion, view the propagation of a light pulse from one point to another. This can be approached quite generally from the Lorentz transformations, Eq. (11). Referring to Fig. 6a, the event ϵ_1 is the transmission of a light pulse, viewed in the S' frame, from the point ($x' = -L' \cos\theta'$, $y' = L' \sin\theta'$, $z' = 0$) at the time $ct' = 0$. The light pulse propagates with speed c to the origin of S', at $x' = y' = z' = 0$, where it

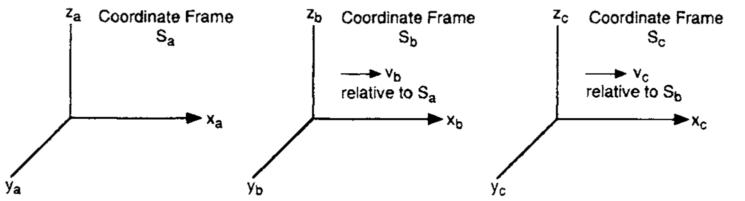


Fig. 5 Three coordinate frames S_a, S_b, S_c where S_b moves at velocity v_b with respect to S_a, and S_c moves at velocity v_c relative to S_b, both in the x direction.

Table 6 General Lorentz velocity addition transformation

Consider a particle in an inertial frame S' that moves at velocity $u'_x = \dfrac{dx'}{dt'}$, $u'_y = \dfrac{dy'}{dt'}$, $u'_z = \dfrac{dz'}{dt'}$, i.e., $\boldsymbol{u}' = (u'_x, u'_y, u'_z)$ with respect to S. Frame S' again moves at velocity v in the x direction with respect to S. The velocity of that particle in the S frame is then obtained from the following differentials:

$$
\begin{aligned}
dx &= \gamma(dx' + v dt') &&= \gamma(u'_x dt' + v dt') = \gamma dt'(u'_x + v)\\
dy &= dy' &&= u'_y dt'\\
dz &= dz' &&= u'_z dt'\\
dt &= \gamma(dt' + v dx'/c^2) &&= \gamma(dt' + v u'_x dt'/c^2) = \gamma dt'(1 + u'_x v/c^2)
\end{aligned}
$$

Thus, the velocity as measured in the S frame is $\boldsymbol{u} = (u_x, u_y, u_z)$ is as follows:

$$u_x = \frac{dx}{dt} = \frac{u'_x + v}{(1 + u'_x v)}$$

$$u_y = \frac{dy}{dt} = u'_y/\gamma(1 + u'_x v/c^2)$$

$$u_z = \frac{dz}{dt} = u'_z/\gamma(1 + u'_x v/c^2)$$

This expression generalizes the velocity addition formula.

arrives at the time $t' = L'/c$. The coordinate intervals along the path in frame S' are $c\Delta t' = L'$, $\Delta x' = L' \cos \theta'$, $\Delta y' = - L' \sin \theta'$, and $\Delta z' = 0$.

Figure 6b shows how the path of the light looks in frame S. Because the Lorentz transformations are linear, coordinate intervals along the light path can be calculated by taking differences in Eq. (11):

$$c\Delta t = \gamma(c\Delta t' + (v/c)\Delta x') \tag{14}$$

$$\Delta x = \gamma(\Delta x' + (v/c)c\Delta t') \tag{15}$$

and $\Delta y = \Delta y'$. The last term $\gamma v \Delta x'/c$ in Eq. (14) is an extra delay caused by the distance traveled by S' during the time of transit.

Consider, for example, the special case $\theta' = 0$. The rod of length L' in S' is laid out parallel to the direction of relative motion and, therefore, to S appears to be shorter, by the Lorentz contraction, so $L = L' \sqrt{1 - \beta^2}$. Light propagates toward the front of the rod, which moves away with speed v. The relative speed with which the light catches up to the front of the rod is $c–v$, so the propagation time in S is as follows:

$$\Delta t = \frac{L'\sqrt{1 - \beta^2}}{c–v} = \frac{\gamma L'(1 - \beta^2)}{c(1 - \beta)} = \gamma \Delta t' \left(1 + \frac{v}{c}\right) \tag{16}$$

On the other hand, this result follows directly from the transformation (14), for substituting $\Delta x' = c\Delta t'$ into Eq. (14):

Table 7 Four-velocity and its space–time invariant

A *four-velocity vector* in a local inertial frame is defined by its components:

$$\boldsymbol{u} = \{u^0, u^1, u^2, u^3\} = \left\{\frac{dx^0}{d\tau}, \frac{dx^1}{d\tau}, \frac{dx^2}{d\tau}, \frac{dx^3}{d\tau}\right\} = \left\{\frac{c\,dt}{d\tau}, \frac{dx}{d\tau}, \frac{dy}{d\tau}, \frac{dz}{d\tau}\right\}$$

where $d\tau$ is the proper time on a hypothetical clock carried along with the particle. $d\tau$ is a scalar invariant, $d\tau = |ds|/c$ along the particle's path. If the particle velocity has magnitude v_p, then

$$d\tau = \sqrt{1 - v_P^2/c^2}\,dt = dt/\gamma(v_P)$$

where

$$v_P^2 = (v_P^1)^2 + (v_P^2)^2 + (v_P^3)^2 \quad \text{and} \quad \gamma(v_P) = (1 - v_P^2/c^2)^{-1/2}$$

In another frame S', the four-velocity vector will have components $u'^\alpha = dx'^\alpha/d\tau$, and under a Lorentz transformation such as Eq. (11) the four-velocity components transform as a contravariant vector (see Table 4):

$$u^\beta = \frac{\partial x^\beta}{\partial x'^\alpha}u'^\alpha \quad \text{or} \quad u^0 = \gamma(v)(u'^0 + vu'^1/c),\ u^1 = \gamma[v](u'^1 + vu'^0/c),\ u^2 = u'^2,\ u^3 = u'^3$$

where $\gamma(v)$ is the time dilation factor of the Lorentz transformation.

Operating on two four-velocity vectors with the metric tensor yields a scalar; e.g.,

$$\boldsymbol{g}(\boldsymbol{u}, \boldsymbol{v}) = g_{\alpha\beta}u^\alpha u^\beta = -u^0v^0 + u^1v^1 + u^2v^2 + u^3v^3$$

where for any inertial frame with Cartesian coordinates the metric tensor $g_{\alpha\beta} = \eta_{\alpha\beta}$, where $\eta_{\alpha\beta}$ has been defined earlier. The first term with a minus sign arising from η_{00} corresponds to the time variable term in the invariant interval, (6), while the other three terms correspond to the spatial variables. The quantity $\boldsymbol{g}(\boldsymbol{u}, \boldsymbol{u})$ is invariant under any Lorentz transformation:

$$\boldsymbol{g}(\boldsymbol{u}, \boldsymbol{u}) = -(u^0)^2 + (u^1)^2 + (u^2)^2 + (u^3)^2 = -\gamma[v_P]^2c^2 + \gamma[v_P]^2v_P^2 = -c^2$$

because $\gamma(v_P) = (1 - v_P^2/c^2)^{-1/2}$.

$$c\Delta t = \gamma(c\Delta t' + (v/c)c\Delta t') = \gamma c\Delta t'(1 + v/c) \tag{17}$$

In the more general case where the length L' is oriented at a nonzero angle to the x,x'-axes, then $\Delta x' = L' \cos\theta'$, and the Lorentz transformation gives the following:

$$c\Delta t = \gamma c\Delta t' + \gamma(v/c)L' \cos\theta' \tag{18}$$

and the extra time delay is $\gamma vL' \cos\theta'/c$. That is, the correction is proportional to the projection of the vector $\boldsymbol{x}'$ on the x-axis. This correction term can, therefore, be written $\gamma\boldsymbol{v} \cdot \boldsymbol{x}'/c$. The correction survives even if second-order effects such as Lorentz contraction are neglected. As shown later, an analogous effect occurs with movement of the Earth's surface where the delay corresponds to a projection of the light pulse position on the equatorial plane.

6. *Drag Effect–Light Velocity in a Moving Dielectric*

Einstein's velocity addition formula also applies to the computation of the velocity of light traveling in a moving dielectric. Consider a dielectric with a

Table 8 Four-momentum for particles and photons

The *four-momentum* of a particle with rest mass m and velocity $\boldsymbol{v} = \{v^1, v^2, v^3\}$ and $v = |\boldsymbol{v}|$ is equal to

$$\boldsymbol{p} = \{p^0, p^1, p^2, p^3\} = \{E/c, p^1, p^2, p^3\} = m\boldsymbol{u}$$

where $\boldsymbol{u}$ is the four-velocity and where $p^0 = E/c = m\gamma c = mc/\sqrt{1 - v^2/c^2} \approx mc + 1/2\ mc(v/c)^2$ for small v, and where $p^1 = m\gamma v^1$, $p^2 = m\gamma v^2$, $p^3 = m\gamma v^3$.

For the four-momentum vector $\boldsymbol{p}$, the quantity $\boldsymbol{g}(\boldsymbol{p}, \boldsymbol{p})$ is invariant under Lorentz transformation:

$$\mathbf{g}(\mathbf{p}, \mathbf{p}) = -E^2/c^2 + (p^1)^2 + (p^2)^2 + (p^3)^2 = \boldsymbol{g}(m\boldsymbol{u}, m\boldsymbol{u}) = m^2\boldsymbol{g}(\boldsymbol{u}, \boldsymbol{u}) = -m^2c^2$$

and is proportional to the squared mass of the particle.

A photon has zero rest mass. Along the path of a photon $ds^2 = 0$; there is no frame S for which a photon is at rest. A photon carries energy and momentum described by its four-momentum vector $\boldsymbol{p} = \{E/c, p^1, p^2, p^3\}$. For a photon $\boldsymbol{g}(\boldsymbol{p}, \boldsymbol{p}) = 0$ and $E^2/c^2 = (p^1)^2 + (p^2)^2 + (p^3)^2$. Energy E and frequency f are related by $E = hf$, where $h = 6.6256 \times 10^{-34}$ Js is Planck's constant. More generally, if the wave number four-vector for photons is $\boldsymbol{k} = \{k^0, k^1, k^2, k^3\} = \{2\pi f/c, k^1, k^2, k^3\}$ then the four-momentum and wave number four-vector are related by $\boldsymbol{p} = h\boldsymbol{k}/2\pi$. Then the Lorentz transformation describes the Doppler shift formula for photons. For example, if in frame S', for which the Lorentz transformations to frame S are given by Eq. (11), a photon has frequency f' and propagates parallel to the x'-axis, then the x'-component of the wave vector is $k'^1 = 2\pi f'/c$, and the frequency f of the photon in S is given by the Lorentz transformation:

$$2\pi f/c = \gamma(2\pi f'/c + \beta 2\pi f'/c) \quad \text{or} \quad f = \gamma(1 + \beta)f' = \sqrt{\frac{1+\beta}{1-\beta}}f'$$

Table 9 Key definitions for GPS relativistic effects

x^α	four-vector space–time coordinates $\{x^0, x^1, x^2, x^3\}$, where x^0 is time expressed in meters, $x^0 = ct$
$\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$	basis vectors
$\boldsymbol{x}$	$x^0\boldsymbol{e}_0 + x^1\boldsymbol{e}_1 + x^2\boldsymbol{e}_2 + x^3\boldsymbol{e}_3$ or $\{x^0, x^1, x^2, x^3\}$
$\boldsymbol{u}$	four-vector velocity (space–time) $\{u^0, u^1, u^2, u^3\}$, $\boldsymbol{u} = d\boldsymbol{x}/d\tau$ m/s
$\boldsymbol{a} = d\boldsymbol{u}/d\tau$	four-acceleration in m/s^2
$c\Delta t$	observed time interval, in meters, measured by clocks at rest in the observer's coordinate frame
β	normalized speed v/c
γ	Lorentz time dilation $\gamma = (1 - \beta^2)^{-1/2} = dt/d\tau$ for inertial frames
τ	proper time is the time measured by a perfect clock moving arbitrarily (measured in seconds)
$\boldsymbol{g} = g_{\alpha\beta}$	metric tensor
ds^2	infinitesimal metric line element $ds^2 = \Sigma\Sigma g_{\alpha\beta}dx^\alpha dx^\beta = g_{\alpha\beta}dx^\alpha dx^\beta$
E	energy
$\boldsymbol{p}$	four-momentum of a photon of energy $E = -\boldsymbol{p}\cdot\boldsymbol{u}$
S	inertial coordinate frame
$\mathfrak{S}$	general accelerating, gravitational coordinate frame

Chapter 19

Joint Program Office Test Results

Leonard Kruczynski*
Ashtech, Sunnyvale, California 94088

I. Introduction

FAVORABLE test results can have a tremendous positive influence on the survivability of a program. This is especially true in the early stages of a program and when a program faces stiff budget battles. A 1979 Comptroller General's report to the Congress was titled "The NAVSTAR Global Positioning System: A Program with Many Uncertainties." The Global Positioning System (GPS) was in a particularly vulnerable position. Two prior reviews by the Comptroller General had made recommendations about the program that, although agreed to by the Department of Defense, had not been implemented. The majority of the questions revolved around funding, schedules, and estimated cost savings. Chapter 4 of the report, however, gave grudging credit to the test results that had been made available. As of October 1978, test data showed that navigation accuracies were better than expected. The favorable test results were one of the bright spots in the Comptroller General's report and undoubtedly contributed to the program's continuation.

In this chapter, we review some of the details of the Joint Program Office (JPO) testing program with an emphasis on the tests conducted at the U.S. Army Yuma Proving Ground (YPG), Arizona. During phase I, Yuma was the site of the majority of tests conducted to support the Concept Validation Phase of GPS. Tests were also conducted by the Navy near San Diego. In phase II, testing expanded worldwide, but YPG remained the heart of JPO testing. U.S. military testers took the GPS receivers to many sites around the world. NATO personnel, an integral part of the JPO, tested GPS equipment in their own countries. Agencies such as the Department of Transportation also conducted tests in support of phase II. The phase III approval carried requirements for equipment modification. Tests continued into phase III at Yuma to validate the modifications that had been made to the equipment.

When not in conflict with GPS user equipment (UE) testing, the JPO supported tests at Yuma that expanded the horizons of GPS. Real-time differential tests were

* Ph.D.

conducted as early as 1979. The experiences using the ground transmitters (GT) at Yuma provide much information to those who would use pseudolites. Detailed and rigorous testing on the F-4 weapon delivery platform provided proof of the potential weapon delivery accuracy. The results of the F-4 tests became a benchmark for future weapons delivery tests. When selective availability (SA) and antispoofing (AS) were tested early in the program, the YPG GTs were used because Block I satellites cannot implement clock dither or AS. The JPO-developed test capability proved to be a highly valuable resource in the development of GPS.

II. U.S. Army Yuma Proving Ground (YPG)

The U.S. Army Yuma Proving Ground (YPG) is the Army's only general purpose proving ground located in desert terrain (Fig. 1). Located about 25 miles from the city of Yuma, Arizona, YPG covers an area of about 1,400 square miles or 870,000 acres. Originally activated in 1943, the Yuma Test Branch's mission was to test bridges, boats, vehicles, and well-drilling equipment. In 1962, what was then Yuma Test Station was assigned to the U.S. Army Material Development Command and was placed under the immediate control of the U.S. Army Test and Evaluation Command (TECOM). The name was changed to Yuma Proving Ground in 1963 and finally to U.S. Army Yuma Proving Ground in 1973. In 1974, YPG was designated a DoD major test facility partly because of the selection of YPG as the primary test site for GPS user equipment.

Yuma Proving Ground's missions include testing tube artillery systems, aircraft armament systems, air delivery systems, and air movable equipment. The missions also include planning, conduct, and reporting of the results of desert environmental

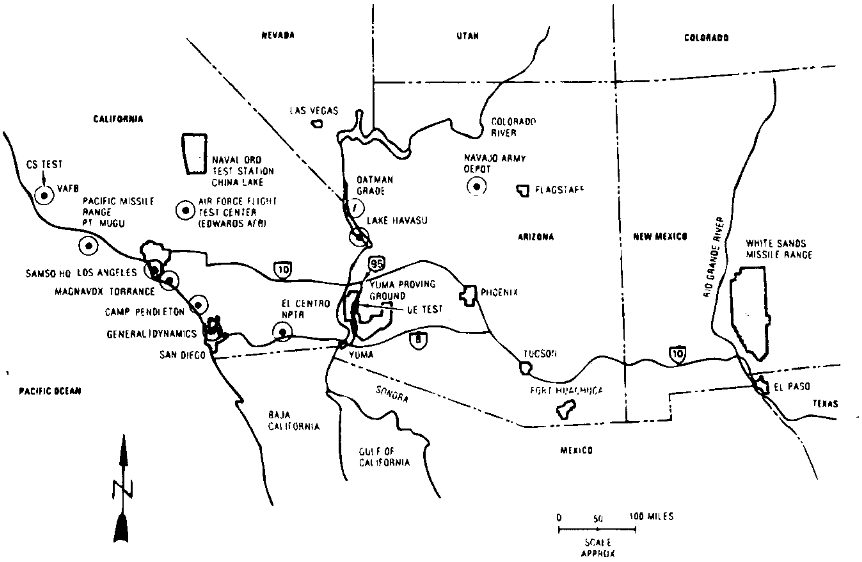


Fig. 1 GPS test locations.

tests of all classes of materiel. The M-1 Abrams tank was a heavy user of the range during the time of GPS testing.

III. Reasons for Selection of Yuma Proving Ground

Yuma Proving Ground has a climate characterized by clear skies, with an average visibility of 25 miles and only three and one-half inches of rain per year. With the extensive real estate and restricted air space, uninterrupted testing can be almost guaranteed. YPG had embarked on a program to use lasers to track artillery shells. The range proposed to use the laser trackers to determine truth trajectories for GPS tests.

IV. Lasers

A total of six lasers were eventually installed at YPG. Figure 2 shows the location of the six lasers and the area of the range covered by at least two lasers. Four lasers were in the southern half of the range and two in the northern half. The tracking lasers are Neodymium YAG lasers built by Sylvania Electronics. The system is designated PATS for precision automated tracking system. The specified radiated power is 50 mJ per pulse, which is automatically attenuated to eye safe levels by measuring the return signal level. Beam divergence, as used in eye safety tests, is less than 2 mrad. Eye protection is required when closer than 2 km to an operating laser.

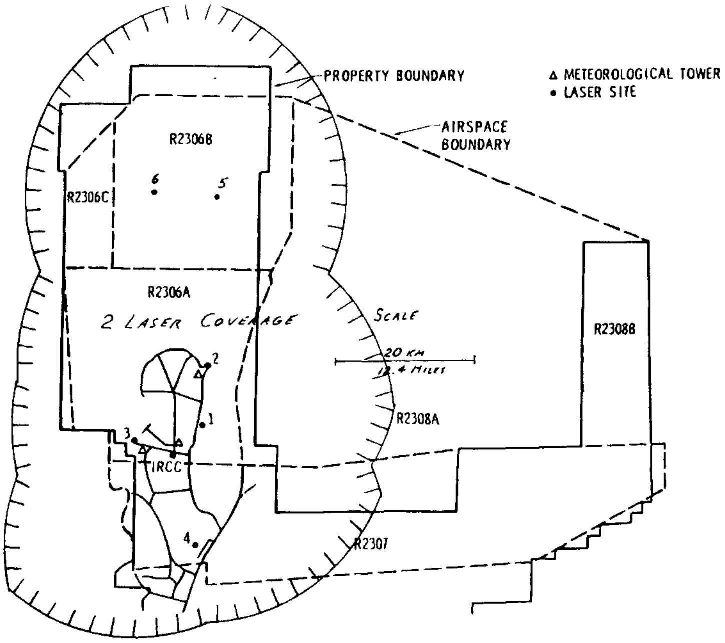


Fig. 2 Yuma Proving Ground area with laser coverage.

Each tracker takes range, azimuth, and elevation measurements at a 100-Hz rate. Specification accuracy for the range measurements is 0.5 m rms from 200 m to 10 km and 1 m rms from 10 km to 30 km. Azimuth and elevation measurement accuracy is 0.1 mrad in each axis. Every fifth measurement is transmitted from the laser site to the Range Operations Center (ROC). No smoothing or editing is performed on the measurements at the laser except to aid in target tracking by the laser. The transmitted measurement is selected only to synchronize the measurements so that a measurement is available on an even second as given by IRIG time. At the laser site, the measurements are recorded at a 100-Hz rate. This field tape is used only for fault isolation. No processing is done on the 100-Hz measurements.

Prior to and immediately following each mission, the laser takes measurements to a series of short targets circling the laser. Typically, there are eight short range targets roughly 1 km from the laser. The short range calibrations are processed by the ROC to provide corrections to tilt (two terms), azimuth, elevation, and range terms.

Each laser also has eight long-range targets at ranges from 1 km to 30 km. The primary use for the long-range calibrations is to determine range bias. Long-range calibrations normally are accomplished monthly or upon request.

"Dump cals" is a third type of calibration for the lasers. To do a dump cal, the laser housing is rotated in elevation through 180 deg, and the short-range targets are tracked. Dump cals are accomplished after maintenance and upon request.

V. Range Space

The Cibola Range covers a ground area approximately 50 km × 100 km and extends to 25 km altitude. Most of the tests were conducted in the southern half of the range. A hilly ridge separates the southern half from the northern half. Two lasers were installed in North Cibola to support aircraft testing that required the entire range. The difficult logistics involved in supporting North Cibola resulted in low utilization of North Cibola. However South Cibola was more than adequate for the majority of testing that had to be accomplished.

Ground vehicle testing was conducted on the dynamometer course. Vehicles on the dynamometer course were tracked by a single laser. Also, the location of the GTs on hills allowed the vehicles to track pseudolite signals when necessary.

VI. Joint Program Office Operating Location

The official presence of the JPO at YPG was the operating location, OL-AA. The operating location normally consisted of Air Force officers and enlisted personnel, civilians assigned by YPG Materiel Test Directorate, military assigned by YPG, Aerospace Corporation personnel, and other agency representation depending on needs. Defense Mapping Agency maintained a slot at OL-AA for several years. General Dynamics Electronics Division and later General Dynamics Services Company provided operations and maintenance support for the operating location.

VII. Satellite Constellation for Test Support

GPS orbital parameters had to satisfy several requirements. High-altitude orbits were desired for survivability and to reduce the number of satellites needed for global coverage. A high inclination was needed to provide coverage in polar areas. A 12-h orbit was selected because it allowed each satellite to be visible over the continental U.S. at least once per day. Twenty-four-hour orbits would not have this feature. In addition, satellites in 24-h orbits were farther from the users, thus requiring greater transmitted power. A critical criterion to support testing was to require that a test window, and preferably an identical test window, exist at a single test site each day. For phase I, the GPS constellation was set up to optimize coverage for testing at YPG.

It is easier to explain GPS test window behavior by assuming that the test site is on the equator and that, on the first day of testing, the satellite passes directly overhead. The requirement for constellation repeatability is that on the next day of testing, the satellite must again be directly overhead. We can further assume that the satellite is crossing the equator in the northerly direction, making this point the ascending node of the orbit.

After two orbits, the satellite will again be at the ascending node over the equator. The key is to require that two orbits of the satellite take the same amount of time as the test site takes to move back under the satellite. Satellites move in space independently of the Earth's rotation. To require that the position of the satellite relative to the ground location repeat requires planning. The two orbits must take 23 hours, 56 minutes and 4 seconds because that is how long it takes the Earth's rotation to bring the test site back to the satellite's ascending node. If this timing is not satisfied, then the location of the satellite with respect to the test site will drift daily, eventually resulting in the lack of a test window at the test site. It is this requirement that gives rise to the 4-min regression of the test window each day.

Had the Earth been perfectly round, the test window would repeat not only in character but also time of day each year. However, the Earth is oblate and the satellite is perturbed. The result is an approximate 50-min shift in the test window each year. This phenomenon helps explain why the GPS test window at Yuma moves from night in the summer of one year to daytime in the summer of a later year.

VIII. Control Segment Responsiveness to Testing Needs

The Control Segment that existed for the testing through phase II consisted of a Master Control Station (MCS) at Vandenberg Air Force Base, California, and monitor stations in Alaska, Hawaii, and Guam. Vandenberg was the launch site for the Block I satellites. The configuration of the monitor stations was especially beneficial to testing at Yuma. The satellites have an eastward motion, so they were generally visible to at least one of the monitor stations prior to rising at Yuma. The MCS was able to upload a satellite either prior to or shortly after the satellite rose at Yuma. By avoiding the use of day-old data, the space and control segment errors seen by receivers under test were minimized.

IX. Trajectory Determination at YPG

A. Real-Time Estimate

The workhorse trajectory at YPG is the real-time estimate (RTE), a position and velocity solution obtained from a single laser. The lasers transmit range, azimuth, and elevation measurements at a 20-Hz rate via microwave links to the ROC. In the ROC, the measurements are preprocessed by a PDP 11/35 and sent to a SEL 32/77 for real-time processing. In the SEL, the measurements are processed first at the incoming 20-Hz rate (every 50 ms) and then at a 5-Hz rate (every 200 ms). The 200-ms task takes every fourth 50 ms measurement on the even 0.2 s, as giver. by IRIG time.

In the 50-ms task, the measurements are processed using a QD filter with a memory parameter equal to six. Measurement residuals are calculated by subtracting the measurement predicted using a second-order formula from the actual measurements. The output of the 50-ms task is range, azimuth, elevation, and first and second rates of change of range, azimuth, and elevation.

The 200-ms task selects every fourth 50-ms output. Calibration factors obtained from premission tests are now applied. The measurements are then transformed from laser-centered range, azimuth, and elevation to IRCC-centered Cartesian coordinates. The IRCC is the Inverted Range Control Center and is located in the middle of the southern part of the YPG test range. These measurements are processed in the 200-ms task with a QD filter with a memory parameter of 10. QD processing in the 200-ms task is similar to the 50-ms task, except the state consists of IRCC East, North, and Vertical components instead of range, azimuth, and elevation used in the 50-ms task. Following the 200-ms QD, the software attempts to determine which retroreflector is being tracked.

To provide continuous track at all attitudes, most aircraft have two retroreflectors. A retroreflector is an array of corner cubes that reflect light in the direction from which the light comes. A retro mounted on top of the aircraft is needed to maintain track when the aircraft banks. The bottom-mounted retro is the real-time estimate solution point. If the software determines that the upper retro is being tracked by the laser, then lever arm corrections are made to move the solution to the lower retro point.

Retro identification and lever arm adjustment are normally accomplished using one of two methods. If no inertial data are available, a wind–velocity algorithm is used to estimate the attitude of the aircraft. Look angles from the laser to the target are computed and are used to determine which retro is being tracked. In the second method, attitude data from GPS user equipment can be used to provide aircraft attitude information for the look angle computation.

There is a range of look angles for which there is a high degree of uncertainty as to which retro is being tracked. Should the look angles fall into this range, the validity flag for the laser's trajectory estimate is lowered. The laser validity flag can also be lowered if measurements are outside statistical bounds.

If a laser is declared valid, then the solution is a candidate for the RTE. The RTE is the truth trajectory for evaluation of GPS UE. Typically, it is the solution determined by the laser closest to the target if the laser's solution has been declared valid. There is an optional composite laser solution that can be used as the RTE. The composite is a weighted average of the valid laser solutions.

Weighting is a function of range to the target. The closer the laser is to the target, the higher the weighting given the laser solution. There is a maximum gain that can be given any laser.

B. Best Estimate of Trajectory

In critical cases, it is possible to compute a best estimate of trajectory (BET), which considers multiple laser inputs. When requested, BETs are computed postprocessed. The quality of the RTE was sufficient for most tests and BETs were computed sparingly.

C. Validation of Truth Trajectory Accuracy

Performance evaluation of GPS UE must be made using a trajectory estimate that is more accurate than the GPS equipment being tested. Laser tracker specifications indicate that the YPG capabilities were within the requirements to evaluate GPS equipment. To support YPG's ongoing program to improve laser accuracy, validation tests were conducted using the Texas Instruments five-channel phase I High Dynamics User Equipment (HDUE). The HDUE was integrated with a Honeywell inertial measuring unit and installed on a Convair 880. The equipment was operated in a differential navigation mode to remove Space and Control Segment errors. The conclusions reached during the various validation tests was that the RTE was accurate to within 3 m throughout the range, with better accuracy in areas of optimal laser coverage.

Velocity accuracy of the lasers was shown to be about 0.2–0.3 m/s. This is on the order of GPS accuracy, especially when aided by inertial sensors. Evaluation of GPS velocity accuracy using the RTE-computed velocities had to be considered only in the sense of bounding the errors. The source of the velocity error could have been the laser as well as the UE.

D. Ground Truth

Early in the GPS program, it was evident that the accuracy performance of user equipment would vary from day to day, depending on the behavior of the satellites and the ability of the Control Segment to predict satellite orbital and clock behavior. Ionospheric delays can vary considerably over time. Tropospheric delay, although it is a smaller contributor to navigation error, might also be different from one test time to the next. However, it was incumbent on the government to evaluate the performance of the various manufacturers fairly, regardless of the performance of the Space and Control Segments and regardless of atmospheric activity. The solution was to compare the error in the positions calculated by the UE to a "ground truth."

Ground truth is the absolute error in the solutions computed by a "perfect" receiver. The development specification for the IRCC defined ground truth as "a direct real-time estimate of the expected error in an IRCC located user's computation of position and time." It was expected that all receivers under test on the range would see virtually identical errors. Thus, if ground truth was 15 m East of the surveyed location and if the receiver under test was also 15 m East, then the receiver was deemed to have perfect accuracy performance.

To create ground truths, the Magnavox phase I X-set receiver in the IRCC was used. The IRCC was located in the center of the southern part of the YPG test range (Fig. 2). The topography was flat with virtually no blockages except by IRCC equipment itself, such as the lightning rod. To minimize clock errors, a cesium frequency standard was connected to the X-set.

Ground truths were computed by the IRCC computer using measurements from the X-set. The measurements were adjusted by removing estimated range and satellite clock behavior using the satellite navigation data and the IRCC's true position, by using L_1-L_2 data to remove ionospheric delay, and by using locally collected meteorological data to remove tropospheric delays. Receiver clock offsets and drift were removed by using the clock synchronization navigation solution. If the satellite data are accurate, if L_1-L_2 measurements allow complete removal of ionospheric delays, if the tropospheric delays are correctly modeled, if the receiver clock algorithm is accurate, and if other error sources such as multipath and receiver noise are zero, then the corrected pseudorange will be zero. Based on the evaluation of all error sources, corrected pseudorange measurements greater than 1–2 m were most likely caused by satellite navigation data inaccuracies.

The clock synchronization navigation solution was a two-state Kalman filter. Receiver clock offset and drift were the two states. Corrected pseudoranges were input to the Kalman filter and the output was processed until the filter was in steady state. Values for the clock offset could vary wildly on a point-by-point basis, especially if different satellites were used to provide the input data. However, rapidly changing values for the clock offset are neither an accurate depiction of true clock behavior, nor are they desirable for the calculation of position ground truths. The IRCC assumed the clock filter to be in steady state when the best estimate of receiver time was fairly constant, when the average of the pseudorange residuals from all the satellites was zero, and when the differences between the pseudorange residuals of the satellites tended to be constant.

The ground truth navigation solution was calculated using a five-state Kalman filter. The first three states were the three-dimensional position and the other two states were the receiver clock offset and drift as they would be calculated by a navigating receiver. The Kalman filter was tuned to provide an explicit solution with each new set of measurements but with a small amount of filtering to reduce noise.

Ground truths were presented as errors in solutions. To use the ground truth, the receiver under test had to use the same constellation for positioning as was used by the IRCC. Initially, this was no problem because the constellation was limited. Four satellites were launched from February 1978 to December 1978. The fifth and sixth successful launches occurred in 1980. The seventh satellite was launched in 1983. The ground truth computation was modified to allow for the best five GDOP constellations. With five satellites available, only five four-satellite constellation selections are possible. With six satellites, up to 15 different four-satellite constellations are possible. However, many constellations would have geometries beyond acceptable limits.

The IRCC performed the additional function of controlling the ground transmitters (GTs). Using ground transmitters requires that the phase of each GT be accurately synchronized. The GTs themselves did not have atomic clocks.

Table 1 Phase I major field test objectives

Navigation accuracy	Threat performance
Position accuracy	Jamming resistance
Velocity accuracy	Denial of accuracy
Effects of dynamics on accuracy	
	Environmental effects
Demonstrations of military value	Propeller and rotor modulation
Precision weapon delivery	Foliage attenuation
Landing approach	Multipath rejection
Rendezvous	Ionospheric and tropospheric correction
Photomapping	
Nap-of-Earth operations	System characteristics
Static positions	Satellite and clock and ephemeris accuracy
Combined operations	Acquisition and reacquisition time
Crosscountry	Time transfer
Shipboard operations	Signal levels and signal structure

X. Phase I Tests (1972–1979)

After passing DSARC I in December 1973, GPS entered the concept validation phase of development. The major field test objectives for phase I are identified in Table 1. The user equipment tested at YPG included receivers from Magnavox under contract to General Dynamics, Collins, and Texas Instruments. Table 2 lists the basic characteristics of user equipment tested as part of the concept validation phase. User sets were mounted in specially instrumented pallets on several vehicles, such as those listed in Table 3.

XI. Ground Transmitters

Testing GPS at YPG began on March 15, 1977. At that time, there were no satellites in orbit. Instead, four ground transmitters, or pseudolites, were used to transmit a satellite-like signal to provide a simulated constellation. The locations of the ground transmitters are identified in Fig. 3 as IRGT-21 through IRGT-24. The pseudolites transmitted an L_1 signal modulated with C/A and P-codes. The data transmitted was a unique message because it is impossible to model the dynamics of a receiver on the ground in a Keplerian formulation. The pseudolites did not transmit an L_2 signal because ionospheric delay was not an issue. PRN codes from 33 to 36 inclusive have been reserved for GTs.

The four GTs and the IRCC comprised the inverted range (IR). The IR was used in the early days of GPS testing when no satellites were available and later to supplement inadequate coverage, for special tests such as SA/AS, and for real-time differential tests. When four GTs were used, even the best geometry has poor vertical dilution of precision (VDOP) performance. When operating in a hybrid mode, one satellite and three GTs provide an optimal geometry. As more satellites became available, it was advantageous to minimize the use of GTs, even at the expense of larger dilutions of precision (DOPs).

Table 2 Characteristics of concept validation phase user equipment

User equipment	Frequency	Code	Number of channels	Integrated with inertial	Contractor
X	L_1 and L_2	P and C/A	4	Yes or No	General Dynamics/ Magnavox
Y	L_1 and L_2	P and C/A	1	Yes or No	General Dynamics/ Magnavox
Z	L_1	C/A	1	No	General Dynamics/ Magnavox
General development model (GDM)	L_1 and L_2	P and C/A	5	Yes	Collins
High dynamics user equipment (HDUE)	L_1 and L_2	P and C/A	5	No	Texas Instruments
Manpack (MP)	L_1 and L_2	P and C/A	1	No	General Dynamics/ Magnavox
Manpack/ vehicular equipment (MVUE)	L_1 and L_2	P and C/A	1	No	Texas Instruments

When four pseudolites were used, the DOPs on the range varied significantly and rapidly. Typical horizontal dilutions of precision (HDOPs) at 20,000 ft above the range changed from 2 to 20 in 8 miles. Thus, an aircraft flying at 300 kt would pass through this range of HDOPs in 90 s. VDOPs had several singularities existing over the course. The character of the DOP contours changed when three GTs were used in conjunction with one satellite, but steep gradients were still present.

The IRCC controlled the transmitted power of the GTs during operation. Each pseudolite was equipped with a remotely controlled programmable attenuator. Maintaining proper power levels was necessary to prevent the GT from jamming the receiver. Power levels were adjusted based on the distance between the receiver and the GT. Multiple aircraft tests were not conducted. Power adjustments were especially critical when the receiver lost lock on a satellite because the receiver would switch to C/A code, which was much more susceptible to jamming than the P-code.

When operating in a hybrid mode, the IRCC had to offset the GT clocks and synchronize them to GPS satellite time. The IRCC attempted to keep the GT clocks accurate to within one ns. The IRCC monitored the GT signals and used a uhf radio control system to control the GTs. The pseudolite clocks were driven by temperature stabilized precision oscillators, which were kept in thermal canisters buried six feet under ground at each pseudolite site.

When satellites were used in conjunction with GTs, the aircraft required two antennas; one antenna on the top and a second antenna on the bottom of the aircraft. The upper antenna tracked satellites, while the lower antenna tracked

Table 3 Concept validation phase test vehicles

Vehicle type	GPS set type	Dynamic range
Mobile test van	XU	Static
Man	Manpack, MVUE	Low
Landing craft	XU, YU	Low
Frigate	XU, YU	Low
Armored personnel carrier	Manpack, MVUE	Low
M35 truck	XU, HDUE, Manpack, MVUE	Low to medium
Jeep	Manpack, MVUE	Low to medium
UH-1 helicopter	XU, YU, Manpack, HDUE, MVUE	Medium
C-141 aircraft	XU, XA, YU, YA, GDM, Z, HDUE	High
P3 aircraft	XU, XA, YU, YA, Z	High
F4 aircraft	XA	Very high

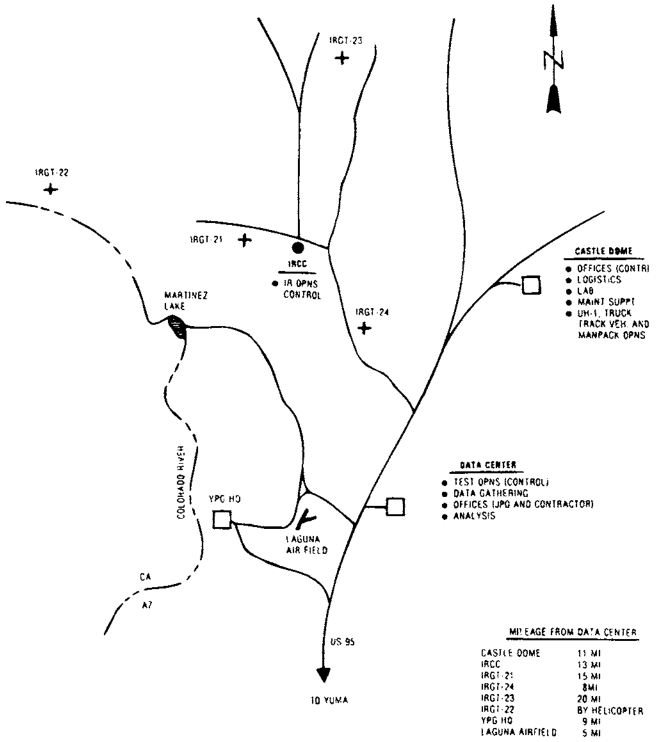


Fig. 3 GPS test area at Yuma Proving Ground showing ground transmitter locations.

the GTs. The phase I UE were designed with two antenna ports. The dual-antenna design also minimized dropouts of GPS satellites during maneuvers. The phase I equipment used inertial attitude data from an external sensor to determine which antenna should be used to track a specific satellite.

The pseudolites on the IR served their assigned purpose of helping to discover and solve user equipment design problems when satellites were not available. Later in the test program, GTs were used to test selective availability and anti-spoofing performance because the Block I satellites did not have SA/AS capabilities. GTs were also used as the data link in real-time differential tests. The experience with GTs has proved valuable to groups investigating the use of pseudolites such as the Range Applications JPO at Eglin Air Force Base.

Phase I test results were excellent. Figure 4 summarizes the performance of the various equipment tested during the conduct of the phase I program. Three-dimensional position errors at the 90% confidence level ranged from about 35 m for the single-frequency Z-set to 15 m for the multichannel, dual-frequency receivers.

Precise statistical values for weapon delivery performance are classified. However, the JPO often presents graphic depictions of the accuracy of weapon delivery. The results from tests have often been displayed relative to the cupola in the center of the Pentagon or relative to a typical suspension bridge.

XII. Navy Testing for Phase I

The U.S. Navy conducted several tests in support of phase I. Most of the tests were conducted in the waters near San Diego. Figure 5 shows the ground track

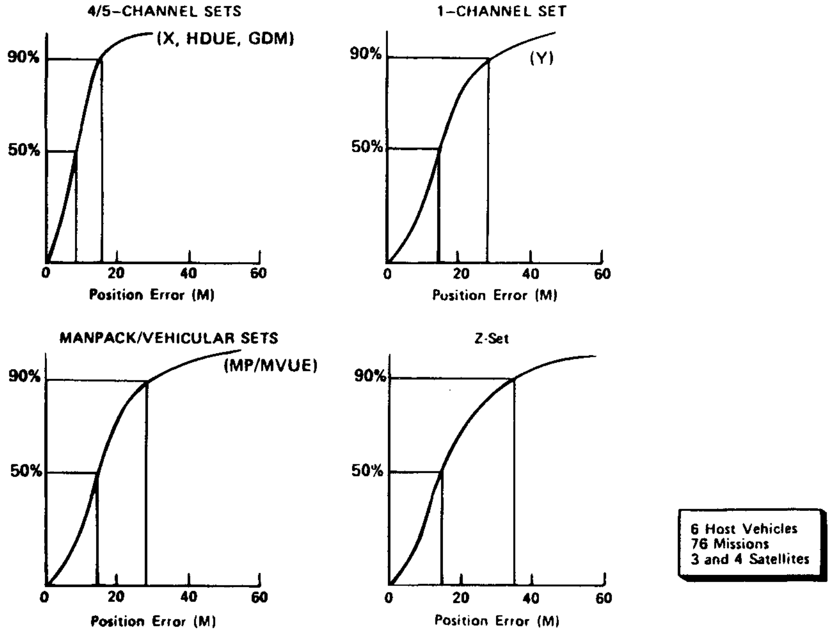
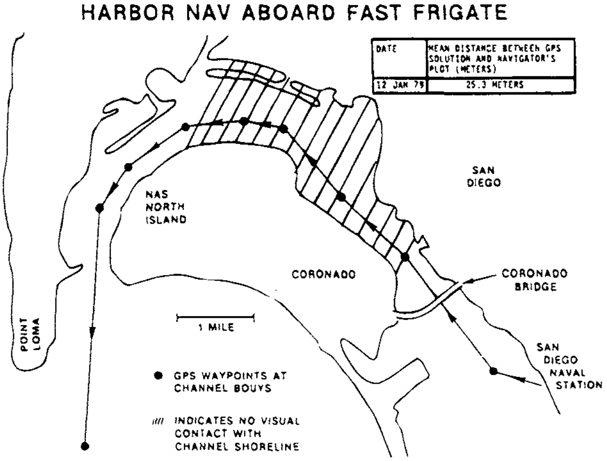


Fig. 4 Phase I results.


HARBOR NAV ABOARD FAST FRIGATE
SAN DIEGO
NAS NORTH ISLAND
CORONADO
CORONADO BRIDGE
SAN DIEGO NAVAL STATION
POINT LOMA
1 MILE
● GPS WAYPOINTS AT CHANNEL BOUYS
//// INDICATES NO VISUAL CONTACT WITH CHANNEL SHORELINE

DATE	MEAN DISTANCE BETWEEN GPS SOLUTION AND NAVIGATOR'S PLOT (METERS)
12 JAN 79	25.3 METERS

Fig. 5 Navigation aboard a fast frigate in San Diego harbor.

of a fast frigate test in San Diego harbor. The hatched area was fogged in during the test. No laser-accuracy truth solution was available for this test. However, the average difference between the navigator's plot and the GPS solution was about 25 m.

Several Navy tests were aimed at identifying and characterizing multipath effects. Dockside testing was accomplished using a Navy landing craft and a fast frigate. Tests were conducted on two subsequent days for each of the craft using P-code receivers. Signal strength fading and pseudorange errors were analyzed to determine multipath effects. Although frequency standard instabilities made the analysis difficult, the tests identified peak-to-peak pseudorange oscillations of 4–5 m caused by multipath.

XIII. Tests Between Phase I and Phase II (1979–1982)

A. Weapons Delivery

Weapons delivery tests continued while YPG awaited the start of phase II. A Navy F-4 was equipped with a Magnavox phase I X-set and integrated with an IMU. Considerable work was done to tune the pilot's display to inform the pilot better about steering information. The software accounted for aircraft attitude, bomb location, ejection velocities, bomb dynamics, ejection velocities, winds estimates, and so forth. The statistical results are classified, but it was not unusual for the inert bomb to hit the stake used to mark the surveyed target point.

B. Differential Tests

Differential GPS was an integral part of range operations from the beginning. Ground truth was based on the concepts behind differential GPS; that is, the removal of errors common to all users in the vicinity. But differential was not limited to simply laying a ground truth plot over a UE error plot. In 1979, a real-time differential system was established to allow the Air Force to conduct flights to validate the performance of the YPG trajectory estimate.

The Texas Instruments HDUE was a five-channel receiver. In normal operation, four channels tracked satellites on L_1 and the fifth channel sequenced through the satellites on L_2. This allowed the receiver to make real-time ionospheric delay measurements.

To modify the HDUE for real-time differential operation, the fifth receiver channel was used to receive the differential corrections transmitted by a GT. The software was modified to allow proper decoding of the GTs data. At first look this may seem to have the effect of removing the capability to compute ionospheric delay. In reality, the differential message transmitted to the HDUE was computed based on a no-ionosphere position computation, in effect, accounting for ionospheric effects.

Tests of the real-time differential HDUE began on December 13, 1979 at YPG on board a UH-1H helicopter. The initial series of tests clearly demonstrated that nondifferential horizontal errors of approximately 20 m are reduced to 5 m or less upon application of the differential correction.

Table 4 is a summary of horizontal and vertical accuracies reported during a series of differential GPS tests in January 1980. Subsequent analysis of the data showed that the processing was flawed and that actual achieved accuracy was better than indicated in the table. The detailed data analysis determined that the processing had not accounted for the lever arm between the retroreflector and the antenna.

The HDUE in differential mode proved to be very accurate. Range validation tests conducted to test the accuracy of the lasers used the HDUE in differential mode. The tester had become the tested.

The early tests with differential mode were accomplished with the corrections transmitted in terms of adjustments to the position solution. As more satellites were launched and it became impossible to count on the receivers using the same constellation, the HDUE differential mode and the IRCC/GT differential message were modified. The IRCC now transmitted corrections to the individual satellite measurements. The HDUE, after receiving and decoding those corrections, applied them as ionospheric delay estimates to the individual pseudoranges. Tests

Table 4 Differential test summaries, m

Date		Mean	Root sum square	Standard deviation	CEP/ probable error
9 Jan 1980	Horizontal	4.2	4.7	2.1	3.8
	Vertical	−5.2	5.5	1.9	5.2
10 Jan 1980	Horizontal	5.0	5.3	1.8	5.0
	Vertical	3.9	4.5	2.2	4.0
11 Jan 1980	Horizontal	3.8	3.8	0.8	3.7
	Vertical	1.4	2.0	1.5	1.5
16 Jan 1980	Horizontal	3.6	3.8	1.5	3.4
	Vertical	2.1	2.6	1.5	2.1

using the measurement correction method began in 1984, with accuracy results similar to previous differential tests. The system again demonstrated its accuracy by identifying an out-of-calibration laser.

XIV. Phase II: Full-Scale Engineering Development Tests (1982–1985)

Phase II, full-scale engineering development testing of GPS UE began in 1982. Two contractors, Rockwell-Collins and Magnavox, had been selected to compete in the phase II program.

In the concept validation phase, it was appropriate and acceptable to tune equipment to specific situations. Satellite selection algorithms and acquisition techniques were not an issue during phase I. Indeed, with the lack of satellites, it would have been difficult to test such requirements. Environmental requirements for the equipment were also secondary considerations in phase I. In phase II, however, the equipment was to be integrated into several host vehicles and the integrations were to be representative of production requirements. Several other applications of GPS were to be tested, especially those related to the integration into complex avionics systems. Phase II test objectives included the following:

1) Verify reaction time, time to first fix, time to subsequent fix, and reacquisition time.

2) Verify static and dynamic position and velocity accuracies.

3) Evaluate IMU/INS alignment and gyro damping capabilities.

4) Determine susceptibility to electronic warfare and nuclear threats.

5) Evaluate precision weapons delivery using GPS bombing mechanization.

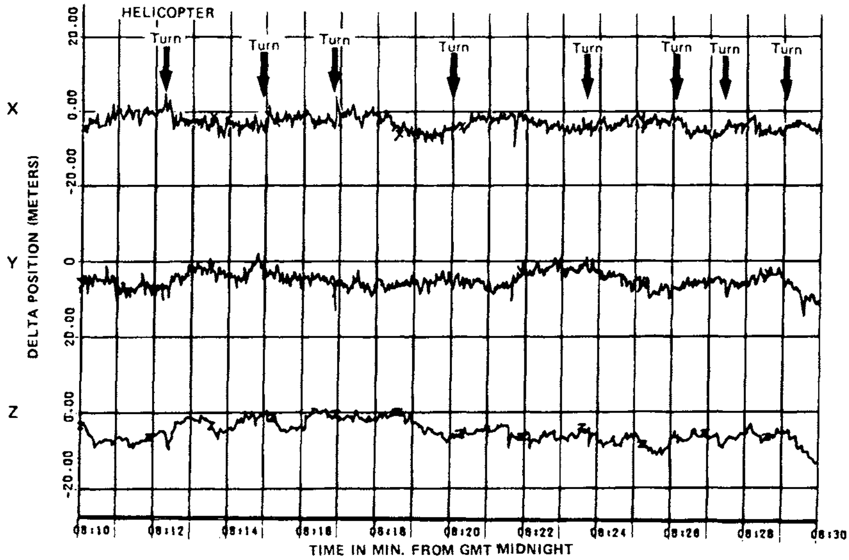


Fig. 6 Example of global positioning system position errors (UH-60 two-channel set).

Table 5 Summary of position accuracy for phase II tests

Vehicle	Spherical error probable, 3D, 50%
A-6	6 m
B-52	9 m
SSN (submarine)	12 m
F-16	13 m
Manpack/vehicular	12 m
UH-60 helicopter	12 m
CV (aircraft carrier)	16 m

6) Determine mission performance improvement for aircraft landing approach, point-to-point and course navigation, rendezvous, ordnance delivery, and carrier-based aircraft alignment and gridlock.

7) Evaluate human factors design.

8) Evaluate reliability, availability, maintainability, and logistics supportability.

9) Identify and track deficiencies and improvements.

10) Evaluate the effects on performance caused by multipath from the surface of the water.

11) Evaluate the adequacy of specified military occupational specialties/skill levels and training.

12) Evaluate the effects on performance resulting from adverse environmental conditions to include RFI/EMI, temperature, humidity, altitude, and vibration.

13) Determine submarine antenna detectability.

14) Evaluate the effects of rotor blade modulation/multipath/masking on performance in an EW environment.

15) Determine ease of chemical, biological, and radiological decontamination.

A sample set of data from phase II testing is shown in Fig. 6. The data represent the East (X), North (Y), and Up (Z) errors in a two-channel GPS set flown on a UH-60 helicopter over the YPG range. A summary of the overall results of the phase II position accuracy tests for various host vehicles is shown in Table 5.

XV. Summary

In this chapter, we have highlighted only some of the massive amount of testing conducted on behalf of the GPS JPO. It will have been over 20 years from the date of the DoD directive that authorized GPS to the date of full operational capability. During that time, technology and politics progressed through several generations and mood swings. Keeping Congress and the public in favor of continued funding of an expensive program is a difficult task. GPS was successful in maintaining the needed funding partly because of the JPO's ability to actually demonstrate capabilities of the system through a series of detailed and rigorous tests. Future users of GPS have a enormous database of information about the details of GPS on which to base decisions about equipment configuration and the appropriateness of application.

Bibliography

GPS User's Overview, Deputy for Space Navigation Systems, Navstar Global Positioning Joint Program Office, YEE-82-009A, Los Angeles, CA, Sept. 1984.

GPS User's Overview, Deputy for Space Navigation Systems, Navstar Global Positioning Joint Program Office, YEE-82-009B, Los Angeles, CA, Sept. 1986.

Henderson, D. W., and J. A. Strada, "NAVSTAR Field Test Results," Paper presented at the Institute of Navigation National Aerospace Symposium, Springfield, VA, March 6–8, 1979.

Kruczynski, L. R., "Truth Trajectory Estimation for Global Positioning User Equipment Tests," Paper presented at the International Telemetering Conference, San Diego, CA, Oct. 24–27, 1983.

Kruczynski, L. R., Abby, D. G., Porter, W. W., and Weston, E. T., "Global Positioning System Differential Navigation Tests at the Yuma Proving Ground."

Mai, R., "The QD Filter in YPG's Real Time Laser Display System," unpublished report by Yuma Proving Ground, June 1978.

"The NAVSTAR Global Positioning System: A Program with Many Uncertainties," Report of the Comptroller General to the Congress of the United States, Jan. 1979.

Teasley, S. P., Hoover, W. M., and Johnson, C. R., "Differential GPS Navigation," Texas Instruments, Inc., TR, 1981.

Weston, E. T., "Inverted Range Navigation Solutions," unpublished General Dynamics Electronics Div. Report, June 1981.

Chapter 20

Interference Effects and Mitigation Techniques

J. J. Spilker Jr.* and F. D. Natali†
Stanford Telecom, Sunnyvale, California 94089

I. Introduction

THE GPS navigation satellite system is designed to serve both military and commercial needs. Because of its military applications, the ability to tolerate significant amounts of interference and jamming was an important consideration in the design of the signal structure. The purpose of this chapter is to describe the effects of interference and to present some design guidelines and mitigation techniques to improve the tolerance to interference of commercial GPS receivers that operate on the C/A signal. The performance of the P(Y)-code signal in the presence of jamming is beyond the scope of this book.

The GPS frequency bands are protected by international and Federal Communication Commission (FCC) frequency assignments. Nonetheless, there is the chance of spurious unintentional interference and even intentional interference. Obviously, any radionavigation system can be disrupted by an interference of sufficiently high power, and the GPS is no exception. Of course, GPS has the obvious limitation that the signal power received has been attenuated by the path loss from the distant satellite. The GPS C/A and P signals are both spread-spectrum signals and, as such, are much less susceptible to interference than a narrow-band signal. For example, a sidetone ranging signal that consists of a pure carrier with tone modulation has a much greater susceptibility to interference. A tone interference could completely disrupt a sidetone- or Doppler-type positioning system if its power simply exceeded that of the desired ranging signal. Interference of even less power could cause significant errors. With the GPS, as described in Chapter 3, this volume, a received tone interference would have to be significantly greater in power than that of the received GPS signal in order to degrade performance appreciably. Thus, the GPS has several advantages over a conventional narrow-band navigation system with respect to purposeful or unintentional interference or disruption for the following reasons:

1) GPS signals, both C/A and P(Y) code, are spread-spectrum signals and permit well-designed receivers to tolerate significantly larger amounts of co-

Copyright © 1995 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Released to AIAA to publish in all forms.

*Ph.D., Chairman of the Board.

† Ph.D., Vice President, Chief Engineer.

channel or adjacent channel interference than can be tolerated by conventional narrowband signals.

2) The GPS position determination is relatively robust, because the system is generally overdetermined; i.e., there are often more satellites visible than the minimum number required for a solution. Furthermore data detection, which has a significantly higher threshold than the tracking functions, is not required continuously.

3) Special receiver design techniques, such as adaptive A/D converters, adaptive frequency notch filters, various types of adaptive delay lock loops (adaptive DLL), and adaptive nulling antennas can further improve receiver performance. An adaptive DLL detects the presence of interference and narrows the tracking loop bandwidth.

4) GPS receivers can be designed to detect interference levels strong enough to disrupt navigation, because the interfering signal must generally be large compared to the relatively stable thermal noise level. Once interference is detected, the receiver can take corrective action.

5) In addition, because the GPS signals are generated on a satellite at a considerable distance from Earth, the satellite signal sources are not easily disrupted by natural disasters. Ground-based transmitters, on the other hand, are sometimes disrupted by violent storms or earthquakes.

The GPS receivers can, by proper design, take maximum advantage of the GPS spread-spectrum signal characteristics. However, beyond a certain interference level, any radionavigation signal can be disrupted. In this situation, the GPS receiver can be designed to detect the presence of the interference and switch to the use of other sensors. This detection process can be similar to that employed in receiver autonomous integrity measurement (RAIM), as discussed in Chapter 5 of the companion volume. However, even in the rare instance when there are an insufficient number of satellites in view to employ RAIM, a high enough level of interference to disrupt GPS navigation can be reliably detected by a properly designed receiver interference detector or by discrete Fourier transform (DFT) processing, and the receiver can alert the operator to take corrective action.

Interference mitigation techniques for GPS receivers include the following:

1) Use of adaptive array, multibeam, or adaptive nulling antennas. These techniques can be categorized as spatial signal processing.

2) Radio frequency/intermediate frequency (rf/IF) filtering to reduce adjacent channel and out-of-band interference.

3) Use of sufficient number of bits/sample and AGC to reduce nonlinear signal suppression effects. Adaptive A/D conversion can reduce constant envelope interference effects.

4) Adaptive filtering in the delay lock loop and phase lock tracking filters.

5) Use of the vector delay lock loop described in an earlier chapter on Signal Tracking Theory (Chapter 7, this volume).

6) Detection of interference followed by an appropriate change in the DLL signal processing and aiding from other sensors, such as inertial measurement units (IMUs).

7) Augmentation of the GPS by ground (pseudolite) and space-borne signal sources.

8) Use of Adaptive frequency notch filters to attenuate narrow-band interference.

A. Possible Sources of Interference

Possible sources of unintentional interference to GPS receivers are shown in Fig. 1. None of these is believed to be a real problem to the GPS, but we should be aware of them, nonetheless. Some of these have at times produced interference to GPS receivers. These include the following[1-3]:

1) Out-of-band interference caused by nearby transmitters coupled with inadequate rf filtering in the GPS receiver.

2) Harmonics or intermodulation products of various ground and airborne transmitters. For example, transmitters at 105.028 MHz, 225.06 MHz, 315.084 MHz, 525.14 MHz, and 787.71 MHz have harmonics (the 15th, 7th, 5th, 3rd, and 2nd respectively) that fall within the GPS L_1 frequency band. Normally these transmitters are sufficiently well filtered to avoid radiating interference.

3) Potential active or passive intermodulation products of signals or local oscillators on the same platform as the GPS navigation receiver or from other nearby platforms. We must ensure that these potential intermodulation products are adequately removed by filtering. For example, we must take care to prevent radiation of local oscillator or intermodulation products from an aircraft transmitter that might couple into a GPS receiver on that same aircraft.

4) Pulsed interference from radar signals in nearby frequency bands that are inadequately filtered.

5) Accidental transmission of signals in the wrong frequency band by experimenters.

B. Frequency Allocation in Adjacent and Subharmonic Bands

We first examine the frequency bands near the GPS frequency band and then other bands that could create harmonics or intermodulation products that fall within the GPS bands if the transmitters are improperly filtered. The frequency bands assigned to satellite radio navigation (GPS, GLONASS) are shown in

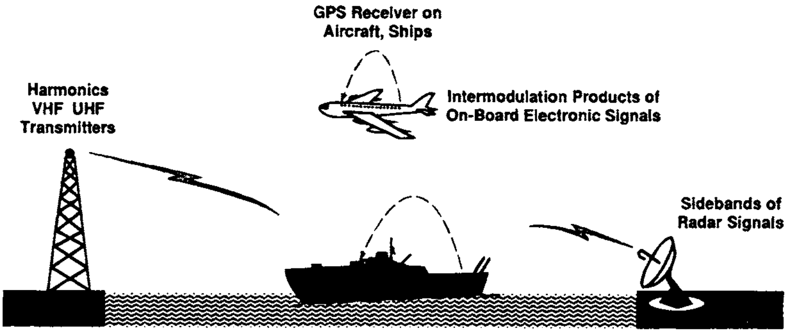


Fig. 1 Possible sources of unintentional interference caused by out-of-band emissions of other signal sources.

Table 1. Other bands in the vicinity of the L_2 signal are presently being considered as a second civil signal. The GPS C/A signal occupies approximately 2.5 MHz of bandwidth centered at 1575.42 MHz (the L_1 frequency).

Nearby bands are assigned for mobile, maritime, aeronautical, satellite communications and satellite-based cellular, as shown in Table 2. The frequency channel just below the GPS L_1 band is a satellite-to-ground link where the power levels received on Earth are expected to be relatively low. On the other hand, the 1610–1626.5 MHz band is a band reserved for hand-held transmitters to transmit to space for satellite-based cellular telephone. Both time division and code division multiple access (TDMA and CDMA) signals are licensed by the FCC for this band. The possible close proximity of these transmitters requires that the GPS receiver have adequate filtering to attenuate these transmissions, which are 34.58 MHz above the GPS L_1 center frequency and 24.58 MHz above the upper (10 MHz) edge of the GPS spectrum.

Another potential source of interference is from satellite communications uplinks at L-band in the 1626.5–1660.5 MHz bands (maritime satrom and aeronautical satrom). If these uplinks are on the same platform as the GPS receiver, there must be enough antenna separation and other isolation to prevent overload of the GPS low-noise amplifier (LNA). In addition one must avoid the effects of active or passive intermodulation products from either satellite communications (SATCOM) multicarriers or from local oscillators in close proximity to the GPS antenna or LNA.

Harmonics of improperly filtered TV channels also can be a potential source of interference in the vicinity of the TV transmission tower. The uhf TV transmission channels are spaced by 6 MHz and extend up to slightly over 800 MHz. Ultrahigh frequency military satellite communication takes place in the 240–400 MHz band with uplinks in the upper portion of the band, primarily in the 290–320 MHz band. Another possible source of interference is a harmonic (12th or 13th) of a VHF civil aviation air-to-ground link in the 118–136 MHz band.

II. Receiver Design for Tolerance to Interference

A. Receiver Systems

A simplified diagram of a GPS receiver that contains both spatial and signal processing for interference suppression is shown in Fig. 2. One or more antennas receive the GPS signal. This signal is rf filtered and downconverted to IF where it is amplified (with AGC) and then sampled and quantized. The quantized signal plus noise is then fed to the parallel code and carrier-tracking loops. All of these

Table 1 Radio navigation frequency bands

Frequency	Bandwidth
1.215–1.240 GHz	25 MHz
1.240–1.260 GHz	20 MHz
1.559–1.610 GHz	51 MHz

Table 2 Frequency bands assigned for mobile satellite communications and satellite-based cellular[a]

Frequency Band	Bandwidth	Usage
1.535–1.559 GHz	24 MHz	Several bands for satellite downlinks to mobile, marine, aeronautical users—space-to-ground
1.610–1.626.5 GHz	16.5 MHz	Uplink based from hand-held units Earth-to-space, satellite-based cellular
782–788 MHz 525 MHz	51 MHz	uhf television—6 MHz channels at 1/2 GPS L_1 frequency uhf television at 1/3 GPS L_1
1626.5–1660.5 MHz	34 MHz	Aeronautical satellite communications uplinks (possible intermodulation products)

[a] Also shown are subharmonic bands that would potentially cause interference if the transmitters are not operating properly.

elements are described in some detail in other chapters. In this chapter, we focus only on the possible effects of an interfering signal, be it broadband or narrowband.

1. Antenna and Adaptive Array Antennas

If only a single GPS antenna is employed, it often is a hemispherical antenna that typically has gain between 0 dB and +6 dB for much of the upper hemisphere. An example antenna pattern is shown in Fig. 3.

Note that at 5 deg elevation angle, the gain for this antenna has dropped to approximately −5 dB. However, as pointed out in the previous chapters, when the elevation angle drops below 5 deg, other effects, such as multipath, physical obstructions, and ionospheric/tropospheric delay uncertainty, may be the dominant error sources. In fact, it may be desirable to attenuate signals received below 5

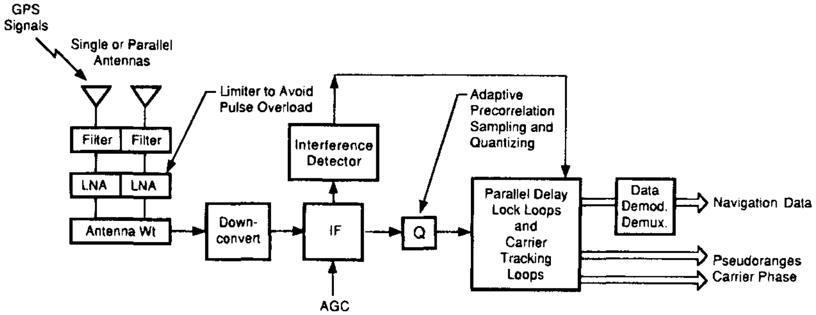

GPS Signals
Single or Parallel Antennas
Filter
Filter
LNA
LNA
Antenna Wt
Limiter to Avoid Pulse Overload
Down-convert
Interference Detector
IF
AGC
Q
Adaptive Precorrelation Sampling and Quantizing
Parallel Delay Lock Loops and Carrier Tracking Loops
Data Demod. Demux.
Navigation Data
Pseudoranges
Carrier Phase

Fig. 2 Simplified diagram of a GPS receiver structure. An antenna array can perform spatial signal processing. An adaptive quantizer can be employed to suppress constant envelope interference. The DLL loop bandwidth can also be adapted. (The LNA is the low noise amplifier.)

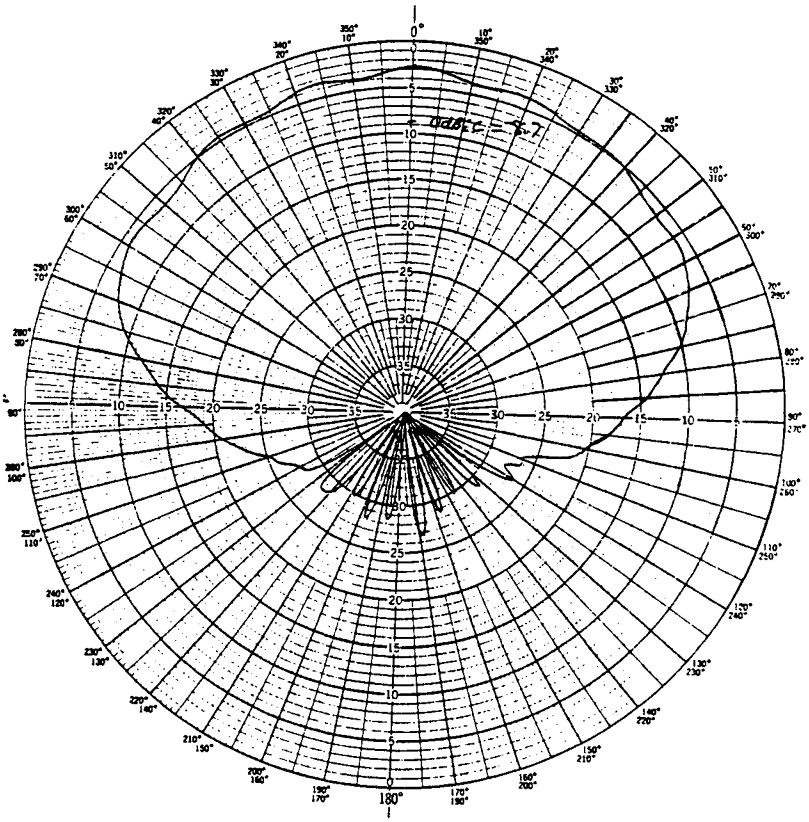


Fig. 3 Typical GPS antenna pattern at L_1 frequency in dB (the gain is 0 dBIC at −8.7 dB on the scale shown) (courtesy of Ball Aerospace).

deg in order to attenuate interfering signals and multipath that might come from the Earth's surface.

Although most GPS receivers employ only a single antenna, some receivers have more than one. For example, two antennas can be employed, one with maximum gain at zenith and another with larger gain at low elevation angles. Satellite receiver channels can be switched between them. Alternatively, a set of antennas can be operated as phased arrays to provide additional antenna gain in the directions to each of the satellites. We can also design an antenna array to null out a single source* of interference by adaptive weighting of multiple antennas, a form of spatial interference suppression. The adaptive nulling of an interfering signal can be performed using an adaptive algorithm very similar to that used in adaptive equalization. Section IV outlines some of the fundamentals of adaptive nulling.

* Multiple nulls can also be created for multiple interference sources as long as there is sufficient separation angle between the GPS satellite and the interference source, and there is a sufficient number of antenna elements.

A multiplicity of antennas can be used in an adaptive array, as shown in Fig. 4.[4,5] The interfering and the satellite signals generally have a significant angular separation if the interference source is on the Earth's surface. Thus, various least mean square and signal-to-noise enhancing adaptive array techniques can be used; e.g., the Applebaum array, to enhance the gain to desired signals and attenuate the interference. Many of these techniques require some knowledge of the signal and the arrival directions, although some adaptive nulling techniques require neither, as discussed in Sec. IV. Constraints can be placed on the adaptive weighting to provide significant gain in the upper hemisphere.

Figure 5 shows an antenna array configured with multiple-beam steering networks. The satellite angle of arrival relative to the user platform must be known to within the beamwidth of the antenna pattern. For fixed ground receiver applications, this information is generally known; for most mobile platforms, however, this operation can be complex.

1. Radio Frequency and Intermediate Frequecy Filtering

Several stages of rf and IF filtering are required. Radio frequency filtering is important to prevent out-of-band interference from causing nonlinear or saturation effects in the LNA and/or intermodulation or image problems in the downconverter mixers. Typically, the user receiver also provides some level of overload protection or limiting prior to the LNA to avoid saturation/burnout of the rf front end caused by high peak power pulses. The rf/IF filter–amplifier combination must be able to reject out-of-band interference and noise, suppress image signal response, prevent intermodulation and spurious effects in the IF amplification and sampling/quantizing operations, while not causing significant group delay variation over the desired signal bandwidth.

The bandwidth of the rf/IF filter for a C/A receiver might be set as wide as 20 MHz in order to get the greatest accuracy in a white thermal noise environment; whereas, this bandwidth could be set as small as 2 MHz (or even 1 MHz) in order to obtain the greater selectivity against out-of-band interference. Figure 6 shows five- and eight-pole Butterworth filter characteristics for bandwidths of

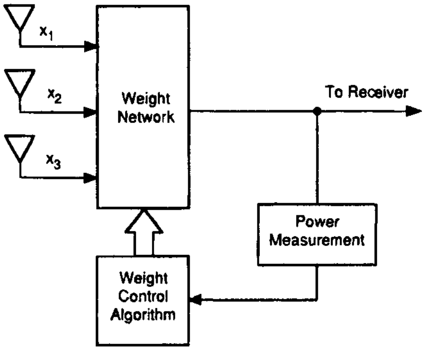


Fig. 4 Adaptive antenna array for nulling interfering signals. The adaptation algorithm can be set to minimize total power because, received power without interference is essentially constant thermal noise.

2 MHz and 20 MHz. Increasing the filter slope selectivity reduces interference effects but increases both the group delay and delay distortion. In summary, although the optimum C/A-code tracking accuracy in thermal noise is obtained by using the full 20 MHz bandwidth, interference rejection and other dynamic range effects may dictate a smaller rf/IF bandwidth.

The AGC sets the signal levels of the receiver so that the performance degradation of the quantizer is minimized. Short duration, high peak power pulses caused by radar sidebands, other pulse interference, or time-gated low duty factor GPS pseudolite transmissions are generally "peak clipped" (limited) or "blanked" (IF disabled for the length of the pulse). The IF or in-phase and quadrature (I,Q) baseband signal is then sampled and quantized. In the next section, the quantizer degradation caused by interference plus thermal noise is analyzed for various numbers of bits per sample and AGC levels. As is shown there, an interfering CW tone with some degree of coherence can cause a 1-bit quantizer to degrade substantially below the performance of that same quantizer with an equal power white thermal noise input. On the other hand, a well-designed multibit A/D converter can actually reduce interference effects for a constant envelope interference.

B. Quantizer Effects in the Presence of Interference

The GPS receiver processing system is assumed to be completely digital; thus, at some point, the received signal plus noise must be sampled and quantized. In this section, the received signal is assumed to contain the desired GPS satellite signal, receiver thermal noise, and an interfering signal. Two forms of the interfering signal are assumed: sinusoidal and Gaussian. A Gaussian interference simply

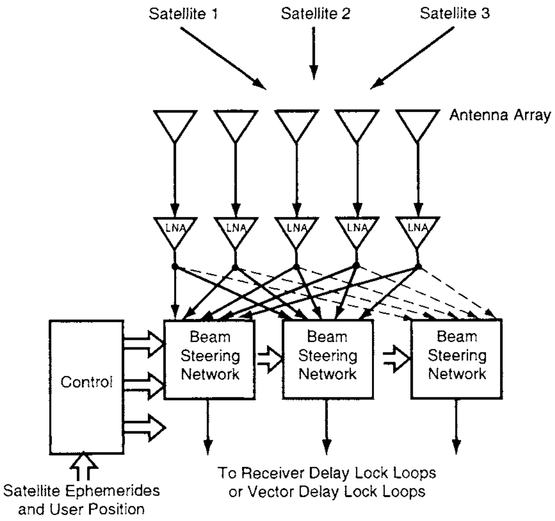


Fig. 5 Multibeam antenna array. This antenna array provides several separate beams, each pointing toward one of the GPS satellites. Approximate knowledge of the satellite angle of arrival is required for this antenna array.

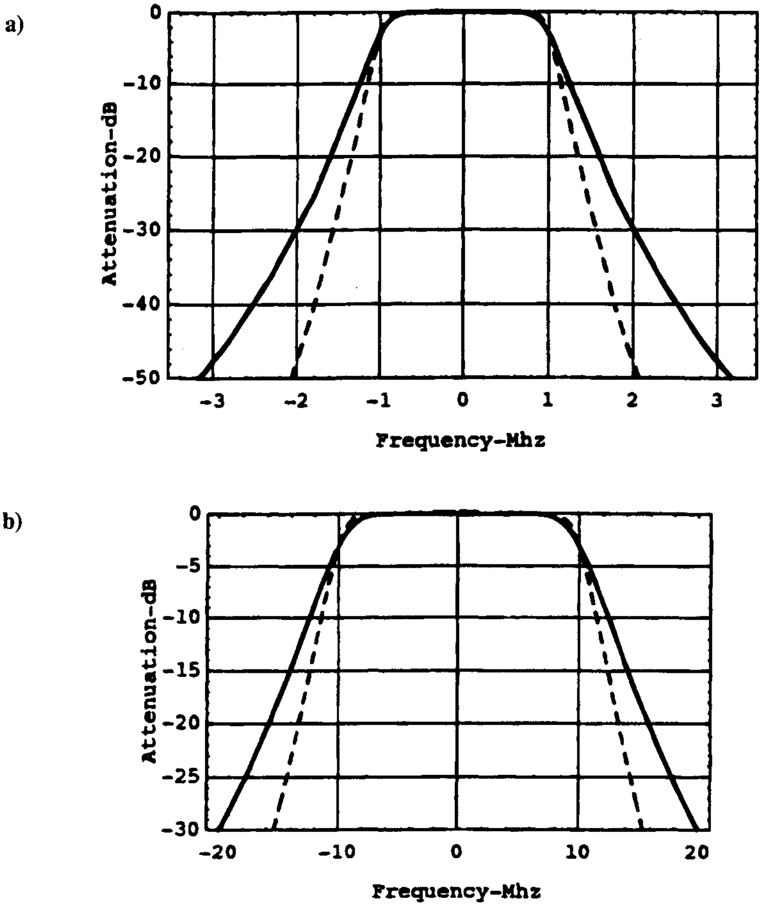


Fig. 6 Example of rf/IF filter frequency response (magnitude) characteristics for five-pole (solid) and eight-pole (dashed) Butterworth filters of 2 MHz and 20 MHz bandwidths.

adds to the Gaussian thermal noise although it may have a different bandwidth. A sinusoidal form of the interference can take on CW, narrowband, and wide bandwidth forms; e.g., an FM signal spread over 100 kHz or 2 MHz. The objective of this section is to analyze the signal-to-noise performance in the output of the quantizer–correlator relative to the input. A sinusoidal interference can have a severe impact on the receiver performance if a 1-bit quantizer is employed or if hard limiting occurs. On the other hand, a well-designed multibit A/D converter can actually suppress a constant envelope interference if the quantizing levels are properly set.

Before we proceed with the detailed discussion of quantizers, it is well to review some results for bandpass limiters. Recall that the GPS signal is a pseudonoise (PN) signal that is biphase modulated on a reference carrier. In a receiver that employs a bandpass limiter,* it is well known that the output signal-to-interference ratio is degraded by 6 dB in the presence of strong sinusoidal interference with significant frequency offset. Frequency offset is defined with respect to the carrier center frequency. That is, the weaker signal is suppressed relative to the strong interference, so it is 6 dB weaker at the output of the bandpass limiter than at the input.[7] If the interference is at exactly the same frequency and in-phase with a desired binary phase-shift keyed (BPSK) PN signal, it can suppress the desired signal by much more than 6 dB and "capture" the receiver. If instead, the interference is a strong Gaussian signal, the suppression of the desired weaker signal by a bandpass hard limiter is $\pi/4$ or -1.05 dB. In this section, we analyze the effects of 1-, $1^1/_2$-, 2-, and 3-bit quantizers on the correlation performance of a receiver where the signal is received in the presence of various coherently modulated constant envelope interfering signals plus white Gaussian noise.

The GPS received signal power is assumed to range from -157 to -160 dBW for the C/A code for a 0 dB gain antenna. If the thermal noise density is -205.2 dBW/Hz, and a 2 MHz or 63 dB-Hz noise bandwidth is used, the total noise power is $P_n = -205.2 + 63 = -142.2$ dBW. The signal-to-noise ratio (SNR) then ranges from $P_s/P_n = \text{SNR} = -14.8$ dB to -17.8 dB for the C/A code. If a bandwidth of 10 MHz (or 70 dB) is used for higher accuracy, the SNR ranges between -21.8 dB to -24.8 dB. In the analyses that follow, a signal-to-thermal-noise ratio of -23 dB is assumed (the quantizer effects are not very sensitive to the SNR ratio in this range). The signal is assumed to be coherent with the receiver local oscillator so that all of the signal power passes through the in-phase channel; whereas, only half of the noise power is passed in the in-phase channel, the other half being in the quadrature channel. Thus, the ratio of signal amplitude level to rms noise level σ for this example is 1/10.

By way of reference it is well known[7] that the SNR degradation caused by a 1-bit quantizer (with high sampling rate) is a factor of $2/\pi$ or -1.96 dB for a signal plus large Gaussian noise (i.e., low SNR).† This degradation decreases to 0.5495 dB for a 2-bit quantizer for the best quantizer spacing relative to the noise level. For a 3-bit quantizer, the degradation decreases to 0.161 dB for optimum spacing.

Figure 7 shows the quantizer characteristic $Q(r)$ for 2- and 3-bit uniform quantizers where the quantizer interval is Δ, and the output levels are ± 1.5, ± 3.5. The quantizers considered here are all uniform quantizers with equal step sizes. Amoroso[8] shows that there is some advantage in interference suppression if nonuniform steps are employed on the output of a 2-bit quantizer, however, that benefit does not occur with thermal noise, and is not considered here.

1. In-Phase and Quadrature (I,Q) Sampling and Quantizing

In Fig. 8, it is assumed that the C/A signal from one satellite has been coherently downconverted to baseband along with Gaussian noise and sinewave interference.

*The bandpass limiter hard limits the envelope of the resultant and passes only the angular information.

†Note that 1-bit quantizing of each of the *I* and *Q* channels is not the same as bandpass limiting wherein a narrowband signal plus noise is hard limited and then bandpass filtered to pass frequency components only in the fundamental frequency zone to the output.

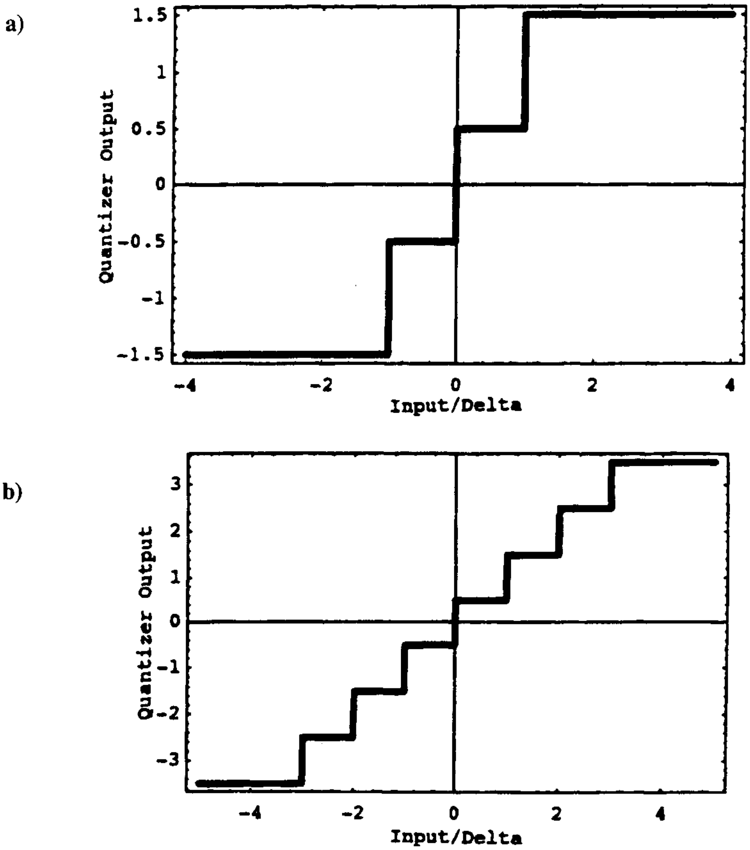


Fig. 7 Quantizer characteristic $Q(r)$ for 2- and 3-bit uniform symmetric quantizers with input quantizer interval Δ.

The interference can be received with or without uncorrelated PN modulation. Assume sufficiently wide bandwidth filters so that filter distortion effects are negligible (or that samples are taken with independent noise samples). The input $r(t)$ to the coherent channel quantizer is then $r(t) = A\ p(t) + K\ q(t) \cos \theta(t) + N_c(t)$, where A is the desired signal amplitude, $p(t)$ is the spreading code, the interference $I(t) = Kq(t) \cos \theta(t)$, and where the PN spreading codes $p(t)$ and $q(t)$ are both assumed to be ± 1. In this model, the interference modulation $q(t)$ may not be at the same rate as the PN code $p(t)$, and, in fact, $q(t)$ may not vary at all. If the interference has no PN modulation and is constant amplitude but offset in frequency by $f = \omega/2\pi$ the interference is $K \cos \omega t$. The bandlimited white noise $N_c(t)$ is assumed to have zero mean, variance σ^2, and Gaussian statistics. The quantizer output is defined as $Q_m[r(t)]$ where Q_m is an m-bit quantizer characteristic. The multiplier (correlator) output is then $c(t) = p(t)\ Q_m[r(t)]$ in the coherent channel.

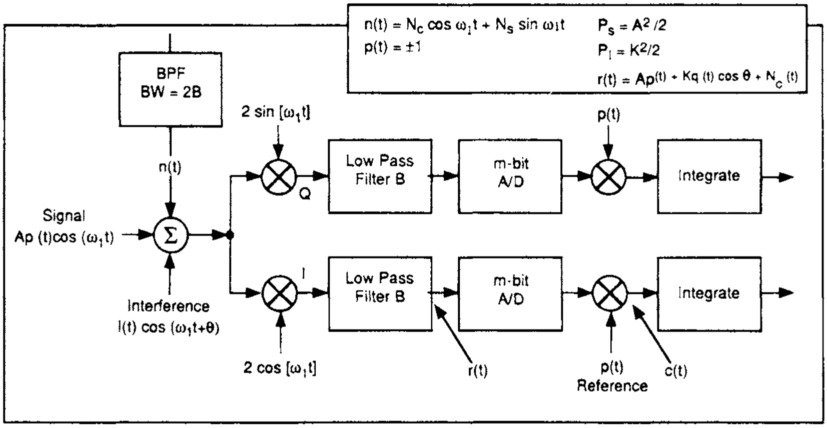


Fig. 8 Digital cross-correlation operation for an input of signal + noise + interference. An *m*-bit A/D converter is used prior to correlation.

The statistics of this multiplier output are examined in the next section. If the signal is sampled at a rate of $2B$ samples per second, the noise samples are independent, because the noise at baseband has bandwidth B. The objective is to compute the ratio of signal power in the correlator output to the variance in the correlator output and compare this ratio with the input signal-to-thermal-noise ratio and the ratio of input signal-to-noise plus interference.

The receiver noise at IF is represented as $n(t) = N_c \cos(\omega_I t + \theta) + N_s \sin(\omega_I t + \theta)$ where $n(t)$ has density N_o (one-sided), and the bandwidth at IF is 2B. The noise power in that bandwidth is $\sigma^2 = N_o 2B$ (see Fig. 9). The power in $n(t)$ is as follows:

$$E[n^2(t)] = E[N_c^2 \cos^2(\omega_i t + \theta) + N_s^2(t)\sin^2(\omega_i t + \theta)]$$

$$= E\left[\frac{N_c^2}{2} + \frac{N_s^2}{2}\right] = E[N_c^2] = \sigma^2 \qquad (1)$$

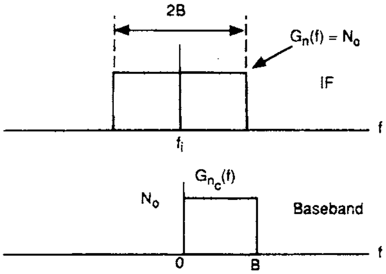


Fig. 9 Noise spectra at rf and at baseband. The baseband spectrum is one-sided.

because $N_c(t)$ and $N_s(t)$ are independent and of equal power, and θ is uniformly distributed; i.e., the variance of $N_c(t)$ is the same as that of $n(t)$. If the signal and noise are coherently mixed with the reference oscillator 2 cos $\omega_i t$, then the output of the in-phase channel for $\theta = 0$ is the product $2 \cos \omega_1 t\,[Ap(t) \cos \omega_1 t + N_c(t) \cos \omega_1 t + N_s(t) \sin \omega_1 t]$. The low-pass component in this in-phase channel is $Ap(t) + N_c(t)$. If the signal power is P_s and the rf noise power is σ^2, then the SNR at the low-pass filter output of the in-phase channel is $A^2/E(N_c^2) = 2\,P_s/\sigma^2$, because $P_s = A^2/2$.

The interfering signal for this example is assumed to have amplitude K and rf power $P_I = K^2/2$. Consider, first, the example where the interference is in-phase with the signal. Other phases and frequencies are considered later for the more general interference model. The total input to the quantizer in the in-phase channel after downconversion is then $r(t) = Ap(t) + I(t) + n(t)$, where $p(t) = \pm 1$ with equal probability, $I(t) = K$, and $n(t)$ has variance σ. The resultant waveform $r(t)$ then has a probability density $\frac{1}{2}$ Normal $(\mu_+, \sigma) + \frac{1}{2}$ Normal (μ_-, σ), where $\mu_+ = K + A$, and $\mu_- = K - A$ are the mean values under these two conditions. The quantizer output $Q[r(t)]$ is then correlated with $p(t)$ to produce the multiplier (correlator) output $c(t)$; i.e.,

$$c(t) = p(t)Q[r(t)] = p(t)Q[Ap(t) + K + n(t)].$$

The correlator-output signal-to-noise ratio is

$$(\mathrm{SNR})_o = E[c(t)]^2/\{E[c^2(t)] - E[c(t)]^2\}.$$

The degradation in quantizer–correlator output SNR vs quantizer input signal-to-thermal-noise ratio* is, then, given by the ratio $R(K,\sigma,A,m,\Delta) = \mathrm{SNR}_o/(A^2/\sigma^2)$, where m is the number of bits in the quantizer, and Δ is the quantizer interval. Define a zero mean Gaussian or normal density function as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x^2/2\sigma^2)} \tag{2}$$

The error function is defined as follows:

$$\mathrm{Erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,\mathrm{d}t, \quad \text{and} \quad \mathrm{Erf}(x_1, x_2) \triangleq \frac{2}{\sqrt{\pi}} \int_{x_1}^{x_2} e^{-t^2}\,\mathrm{d}t \tag{3}$$

and the normal distribution function is

$$\Phi(y) = \int_{-\infty}^{y} p(x)\,\mathrm{d}x$$

Thus, the error function and Gaussian distribution functions are related by the following:

$$\Phi(y) = \frac{1}{2}\,\{\mathrm{Erf}[y/(\sqrt{2}\,\sigma)] + 1\} \tag{4}$$

*Note that the quantizer input SNR is twice the rf input SNR for a coherent receiver. Note, also, that this degradation is not referenced to the ratio of input signal-to-noise plus interference.

and the probability of a random variable y being in the interval (y_1, y_2); namely, $p(y_1 < y < y_2)$, is as follows:

$$\Phi(y_2) - \Phi(y_1) = \frac{1}{2}\,\mathrm{Erf}[y_1/(\sqrt{2}\,\sigma), \quad y_2/(\sqrt{2}\,\sigma)] \tag{5}$$

Also, note that for small x, we have the following:

$$\mathrm{Erf}(x) \cong \frac{2}{\sqrt{\pi}}\left[x - \frac{x^3}{3} + 0(x^4)\right] \quad \text{for} \quad x << 1 \tag{6}$$

The expected value of the correlator output is the mean value of $Q[r]p$. For a 2-bit quantizer with output levels ± 1/2, ± 3/2, where $p(t) = \pm 1$ with equal probability and all levels normalized to $A = 1$ the expected value of the correlator output is† (see Appendix):

$$\begin{aligned} E(c) = {} & \frac{1}{2}\left[P\left(Q = +\frac{1}{2} \mid p = 1\right) - P\left(Q = \frac{1}{2} \mid p = -1\right)\right]\frac{1}{2} \\ & + \frac{3}{2}\left[P\left(Q = +\frac{3}{2} \mid p = 1\right) - P\left(Q = \frac{3}{2} p \mid = -1\right)\right]\frac{1}{2} \\ & - \frac{1}{2}\left[P\left(Q = -\frac{1}{2} \mid p = 1\right) - P\left(Q = -\frac{1}{2} \mid p = -1\right)\right]\frac{1}{2} \\ & - \frac{3}{2}\left[P\left(Q = -\frac{3}{2} \mid p = 1\right) - P\left(Q = -\frac{3}{2} \mid p = -1\right)\right]\frac{1}{2} \end{aligned}$$

This expression can be evaluated as follows:

$$\begin{aligned} E(c) = {} & \frac{1}{8}\left\{\mathrm{Erf}\left[\frac{0 - (K+1)}{\sigma\sqrt{2}}, \frac{\Delta - (K+1)}{\sigma\sqrt{2}}\right]\right. \\ & \left. - \mathrm{Erf}\left[\frac{0 - (K-1)}{\sigma\sqrt{2}}, \frac{\Delta - (K-1)}{\sigma\sqrt{2}}\right]\right\} \\ & + \frac{3}{8}\left\{\mathrm{Erf}\left[\frac{\Delta - (K+1)}{\sigma\sqrt{2}}, \infty\right] - \mathrm{Erf}\left[\frac{\Delta - (K-1)}{\sigma\sqrt{2}}, \infty\right]\right\} \\ & - \frac{1}{8}\left\{\mathrm{Erf}\left[\frac{-\Delta - (K+1)}{\sigma\sqrt{2}}, \frac{0 - (K+1)}{\sigma\sqrt{2}}\right]\right. \\ & \left. - \mathrm{Erf}\left[\frac{-\Delta - (K-1)}{\sigma\sqrt{2}}, \frac{0 - (K-1)}{\sigma\sqrt{2}}\right]\right\} \end{aligned}$$

† $P(Q = +1/2 \mid p = 1)$ is the conditional probability that $Q = +1/2$ given that the PN chip $p = 1$.

$$+\frac{3}{8}\left\{\mathrm{Erf}\left[-\infty, \frac{-\Delta-(K+1)}{\sigma\sqrt{2}}\right]-\mathrm{Erf}\left[-\infty, \frac{-\Delta-(K-1)}{\sigma\sqrt{2}}\right]\right\} \tag{7}$$

The variance of c can be computed in a similar manner by averaging the probabilities that c^2 takes on the values 1/4, 9/4. The variance of c for the 2-bit quantizer is, then, as follows:

$$\begin{aligned}\mathrm{var}(c) = {} & \left(\frac{1}{2}\right)^2\left[P\left(Q=+\frac{1}{2}\mid p=1\right)+P\left(Q=+\frac{1}{2}\mid p=1\right)\right]\frac{1}{2}\\ & +\left(\frac{3}{2}\right)^2\left[P\left(Q=+\frac{3}{2}\mid p=1\right)+P\left(Q=+\frac{3}{2}\mid p=-1\right)\right]\frac{1}{2}\\ & +\left(\frac{1}{2}\right)^2\left[P\left(Q=-\frac{1}{2}\mid p=1\right)+P\left(Q=-\frac{1}{2}\mid p=-1\right)\right]\frac{1}{2}\\ & +\left(\frac{3}{2}\right)^2\left[P\left(Q=-\frac{3}{2}\mid p=1\right)-P\left(Q=-\frac{3}{2}\mid p=-1\right)\right]\frac{1}{2}\end{aligned} \tag{8}$$

2. *One-Bit Quantizer—Quantizer–Correlator Degradation*

If the quantizer interval is set at $\Delta = 0$, the 2-bit quantizer of Fig. 7a degenerates to a 1-bit quantizer, and the output is then always $\pm 3/2$. Again consider a received input which consists of a unit amplitude biphase modulated signal plus a coherent interference of fixed amplitude K plus Gaussian noise of standard deviation σ. The probability density of the quantizer input given each of the two signs in the biphase modulation then appears as shown in Fig. 10. For this 1-bit quantizer, a coherently related interference of sufficient size can be shown to capture the receiver totally since the quantizer output then has almost no correlation with the input C/A–PN code $p(t)$.

Figure 11 shows the degradation in quantizer output SNR relative to quantizer input SNR plotted in dB for the signal level $Ap = \pm 1$, a noise level $\sigma = 10$, and interference amplitude K. Note that for small coherent interference amplitude $K = 0.1$, the degradation is 1.9548 dB, very close to the small SNR asymptotic value of $2/\pi$ (1.961 dB) for no interference. As the interference amplitude increases to 10, the degradation increases to approximately 6 dB. At larger interference levels for an interference amplitude level of approximately 25, the degradation rapidly increases to approximately 30 dB. This rapid quantizer–correlator degradation with interference levels not much greater than the receiver thermal noise level makes the use of 1-bit quantizing unacceptable if tolerance to coherent constant envelope interference is important. It can easily be seen that if there is no thermal noise, a coherent interference only slightly larger than the input signal amplitude completely suppresses the signal if a 1-bit quantizer is employed.

It is well known that BPSK PN signals are subject to stronger signal capture effects in a bandpass hard limiter, as shown in Fig. 12. A large interference,

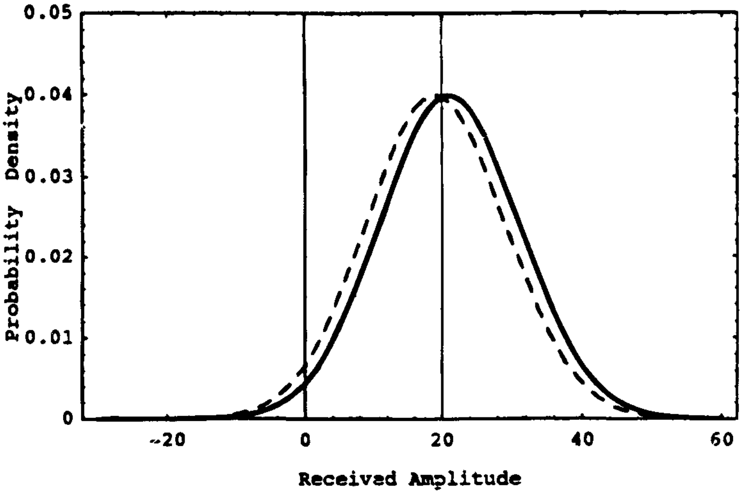


Fig. 10 Probability densities of the resultant of a biphase modulated unit amplitude signal plus coherent interference of level K plus thermal noise with $\sigma = 10$. This resultant signal is the input to a 1-bit quantizer for the two conditions $p = \pm 1$. In this example, $K = 20$, $\sigma = 10$, and $A = 1$. When the interference amplitude $K > \sigma$, the interference level carries the quantizer input to such a level that only very infrequent transitions are made, and the interference begins to capture the receiver totally.

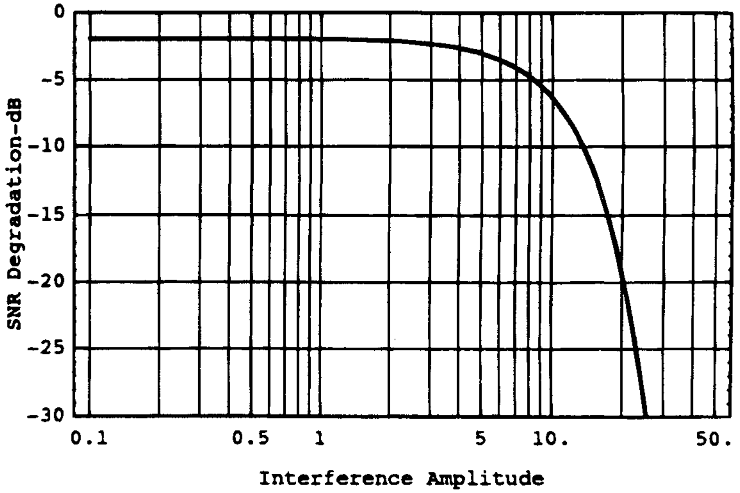


Fig. 11 Degradation in quantizer–correlator output SNR relative to quantizer input SNR caused by a 1-bit quantizer for coherent constant envelope interference of amplitude K relative to the signal amplitude. Gaussian noise is present with $\sigma = 10$. The SNR degradation is defined relative to the input signal to Gaussian noise ratio.

which is momentarily coherent with the signal for that time interval, completely suppresses the signal. Thus, we must avoid bandpass limiting in any GPS receiver with a potential for sinusoidal interference.* These effects with I/Q receiver 1-bit quantization are very similar to those studied by Baer[9] and Aein and Pickholtz.[10] The drastic suppression effects for a 1-bit I,Q quantizer however do not disappear with QPSK PN as they do with bandpass hard-limiter. The interference could be offset by 45 deg from the signal and would affect both I and Q channels equally.

The GPS C/A code, of course, only employs biphase modulation (the quadrature P(Y) code is not generally available to the civil user). Thus, the C/A receiver reference is also biphase modulated. The suppression effect of tone interference on the receiver, however, can be largely avoided through the use of quantizers with two, three, or more bits per sample, and an appropriate AGC, as shown below. Notice also that in the analysis of Fig. 10, the degradation has assumed that the interference is directly in-phase with the signal. If the interference is out of phase by some angle θ, the effective interference amplitude then becomes $K \cos \theta$ instead of K. However, it is clear that, in general, this effect simply means that the interference has to be slightly stronger, and, with θ slowly varying the signal, is suppressed periodically unless a multilevel quantizer is employed.

If the interference is offset in frequency by a large amount compared to the inverse of the averaging time; i.e., the data rate, then the mean and variance of $c(t)$ must be computed by averaging over θ in the $K \cos \theta$ term (with a uniform distribution of θ). This degradation is shown in Fig. 13. Note that for an interference amplitude of 20, the quantizer–correlator degradation is approximately 8.5 dB for the frequency offset interference vs 19 dB for the phase coherent interference, but the degradation is still increasing rapidly with increasing interference, and the 1-bit quantizer performance is unacceptable for this purpose.†

3. *Two-Bit Quantizing—Quantizer–Correlator Performance*

The performance of 2-bit quantizing is examined for PN signal plus noise alone, signal plus noise plus in-phase interference, and finally signal plus noise

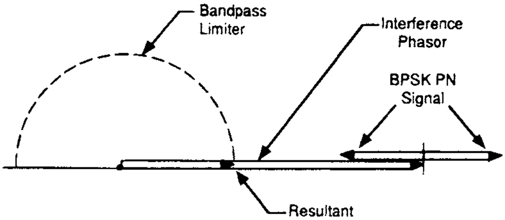


Fig. 12 Sum of BPSK PN signal and a coherent interference of larger amplitude. The resultant phasor is completely independent of the BPSK PN signal for this simple noise-free example, and the signal is completely suppressed. Note that noise is absent in this simple example.

*With QPSK PN, high-level sinusoidal interference would completely suppress only the in-phase term; i.e., half the signal power.

†One-bit quantizers are simple and lead to low-cost receivers, but their susceptibility to interference is a serious flaw for some applications.

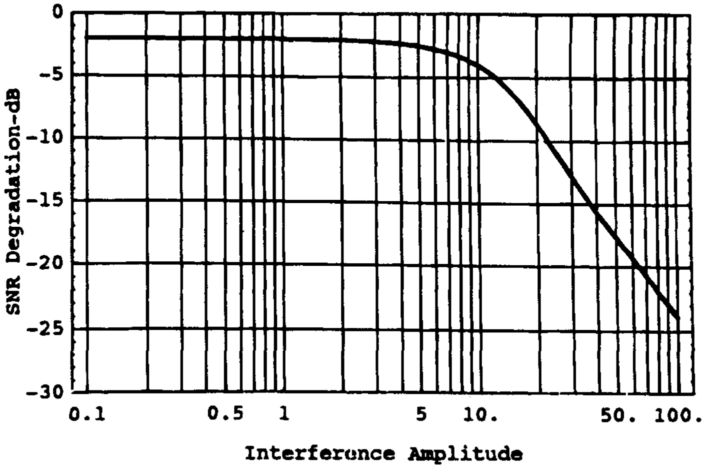


Fig. 13 Quantizer–correlator SNR degradation for a 1-bit quantizer with a frequency offset interfering sine wave $K \sin ax$ of amplitude shown. The signal is received with Gaussian noise ($\sigma = 10$) plus sinusoidal interference. For no interference the SNR degradation is $2/\pi$ or 1.961 dB. For large sinusoidal interference the degradation approaches 6 dB relative to the interference power level.

plus a sinusoidal interference which is offset in frequency from the signal center frequency. Clearly, in the limit for quantizer interval $\Delta = 0$ or ∞, the 2-bit quantizer is either always in saturation, or the received signal plus noise and interference never exceeds the first quantizer magnitude level Δ, and the 2-bit quantizer operates as a 1-bit quantizer. Figure 14 shows the output quantizer–correlator signal-to-noise performance degradation of a 2-bit quantizer relative to the quantizer input signal to thermal noise ratio (or to a receiver with an infinite number of quantizing intervals (linear)). The Gaussian noise has a $\sigma = 10$, and there is no interference. As shown in the figure, the minimum quantizer degradation for this case is 0.5415 dB and occurs at $\Delta = \sigma$. As the value of the input SNR declines to 0; i.e., infinite noise, the degradation increases only slightly to 0.5495 dB. Note that the degradation is still only 1 dB or less for quantizer interval variations of $\Delta = 0.4\ \sigma$ to $1.8\ \sigma$ corresponding to an AGC gain set error of approximately ± 7 dB.

If an interfering signal is present, and it has Gaussian amplitude statistics, the same result of Fig. 13 would also apply to the interfering signal problem. That is, set the AGC levels so that $\Delta = \sqrt{\sigma_n^2 + \sigma_I^2}$ where σ_n^2, σ_I^2 are the noise and interfering power levels, respectively.

Now consider a 2-bit quantizer that is employed with a signal in the presence of coherent interference of amplitude $I_c = K\cos\theta$ (for fixed θ) on the in-phase channel as well as Gaussian noise. The signal is of unit amplitude $Ap = \pm 1$, and the noise has rms value $\sigma = 10$. The quantizer interval is Δ. The quantizer–correlator SNR degradation above that of the quantizer input signal-to-thermal-noise ratio is shown in Fig. 15. Interference levels of $K = 0$, 10, 20, and 40 correspond to quantizer input interference-to-noise ratios of $-\infty$, 0, 6, and 12 dB,

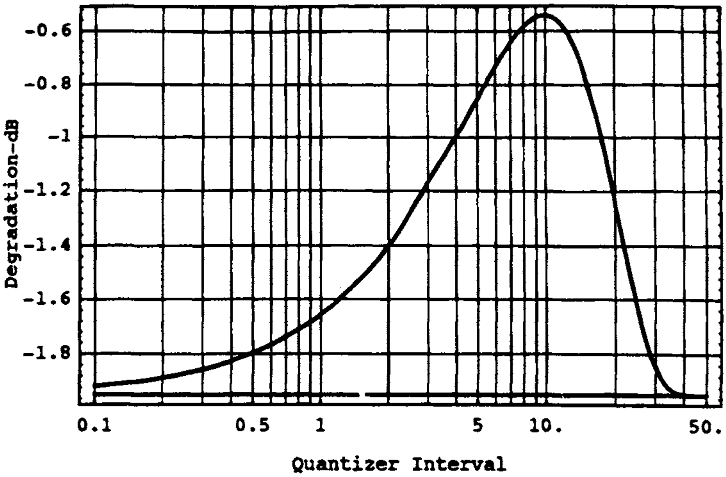


Fig. 14 Quantizer–correlator signal-to-noise degradation in dB for a 2-bit quantizer with normalized quantizer interval Δ for signal plus noise. The degradation is plotted vs quantizer interval Δ. The signal is of amplitude $Ap = \pm 1$, and the Gaussian noise has rms value $\sigma = 10$. There is no interference in this example.

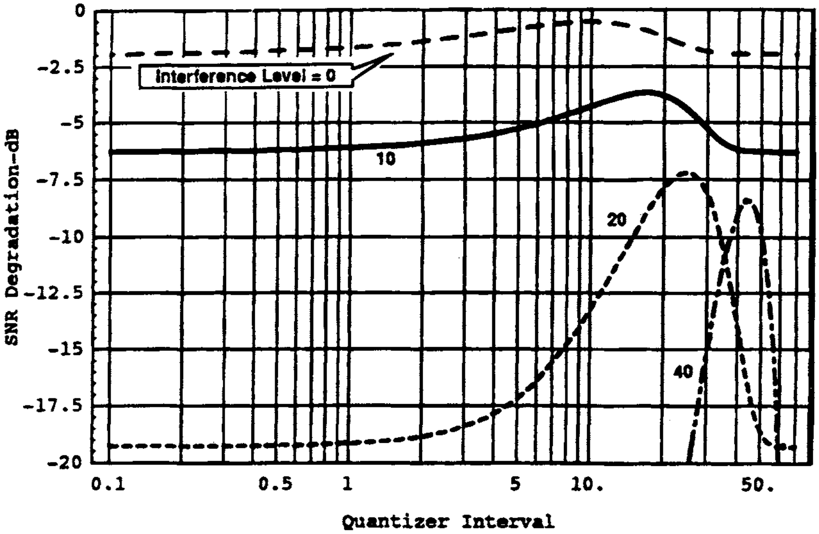


Fig. 15 Quantizer–correlator SNR degradation above that caused by noise alone for the 2-bit quantizer with coherent fixed interference levels of $K = 0$, 10, 20, and 40. If the quantizer was perfectly linear, the added degradation caused by interference would have been 0 dB, 3.0 dB, 7.0 dB, and 12.3 dB, respectively.

respectively. Notice that the range of good quantizer settings grows progressively narrower in percentage as the interference increases. Its width is approximately equal to the standard deviation of the noise. This result implies that the AGC level becomes more critical (when expressed in dB) as the interference level increases.

Note, however, that as the interference increases, the degradation does not increase proportionally with total noise plus interference power, as it would for Gaussian interference. It is clear from these results that the 2-bit quantizer interval Δ (or equivalently the AGC) can be set to attenuate the interference. Amoroso[8] pointed out that it is possible to design the quantizer to be adaptive so as to attenuate a sinusoidal interference. If the quantizer is set at the optimum level for large coherent interference, the input signal plus noise plus interference sits with a peak of the probability density near Δ (recall Fig. 10). Thus, the biphase modulated signal level toggles the magnitude both above and below the quantizer level Δ. In this case, a constant interference in the in-phase channel simply acts as a constant bias, and its effect to a large extent is canceled. Reference 8 showed that this effect can be enhanced by making the quantizer steps nonuniform, in the limit, the quantizer is a $1\frac{1}{2}$-bit null zone quantizer, as shown in Fig. 16. Clearly, in this example, the interference is almost completely canceled for a quantizer setting of $\Delta = K$ for large K.

As is evident from the curves of Figs. 14 and 15, the quantizer setting or AGC can be controlled to make it nearly equally probable to be in the two quantizer amplitude settings. The optimum quantizer setting in Fig. 15 is very close to $\Delta = \Delta_o \triangleq \sqrt{\sigma^2 + K^2}$. For example, at $K = 20$, the value of $\sqrt{10^2 + K^2} = 22.36$.

Note that the fact that the optimum quantizer setting results in equally probable output magnitudes makes it possible to build such an adaptive quantizer relatively easily. We simply have to count the number of occurrences in the two output magnitudes, and adapt the quantizer setting Δ so that they are equal.

Figure 17 shows the performance degradation for different AGC/quantizer settings for the 2-bit quantizer. Note that the optimum quantizer setting for this noise and interference, $\Delta \cong \sqrt{10^2 + K^2}$, gives a degradation that decreases to an

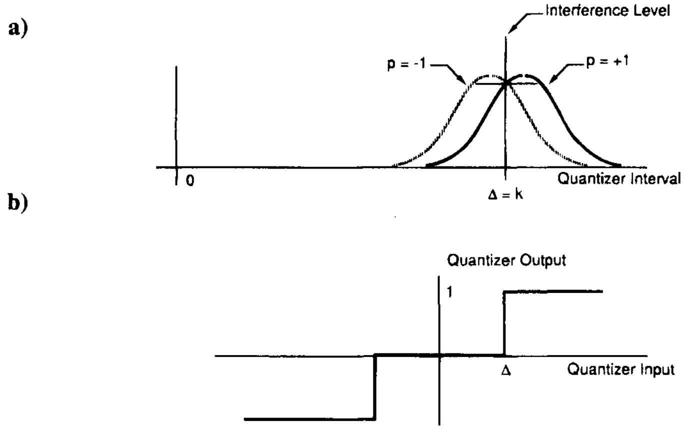


Fig. 16 Null zone quantizer setting with fixed large interference of amplitude K: a) quantizer input probability density; b) $1\frac{1}{2}$-bit, three-level quantizer with a null zone.

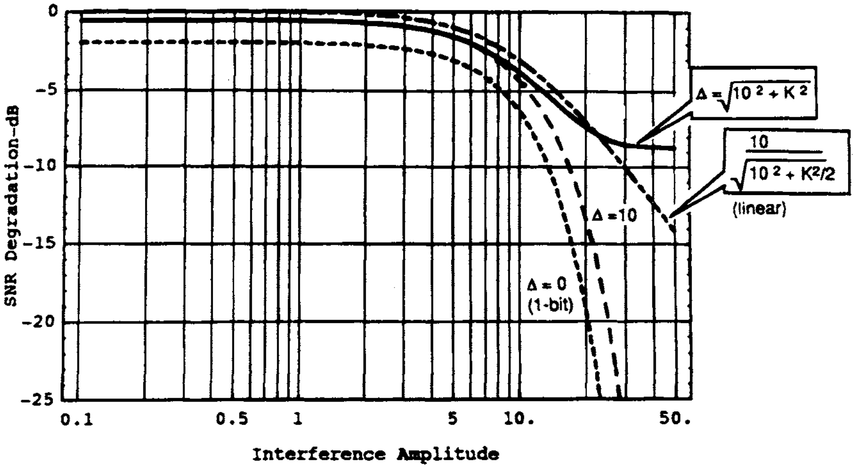


Fig. 17 Signal-to-noise ratio degradation for a 2-bit quantizer for quantizer intervals of $\Delta = \sqrt{10^2 + K^2} = \sigma_{\text{total}}$, $\Delta = 10$, and the 1-bit quantizer $\Delta = 0$. The linear reference degradation $10/\sqrt{10^2 + K^2/2}$ is shown as the alternate dash curve.

asymptote of approximately −9 dB. The reason for this behavior is that, as shown earlier in Fig. 5, a fixed interference level will still toggle the quantizer's output with the sum of random noise and signal. Thus, the interference effect itself is suppressed by the quantizer. Obviously, any time variation in the interference amplitude or phase would cause this interference nulling effect to be reduced, as shown below. As shown in Fig. 15, the minimum in quantizer degradation occurs in the region $\Delta = \Delta_0 \pm 10$. Although this result is for perfectly constant coherent interference, it is shown later that even with a sinusoidal offset interference, there is a definite performance improvement for an optimal quantizer compared to a linear quantizer.

A three-dimensional plot of the performance degradation vs both interference level and quantizer interval is shown in Fig. 18. As can be seen from the plot, there is a clearly defined ridge that optimizes the quantizing interval for any given interference level.

If the interference phase θ varies slowly with time relative to the databit rate, but the AGC maintains the quantizer level at a constant setting $\Delta = \sqrt{10^2 + K^2}$, then the quantizer output degradation varies slowly with phase offset (or time if the phase varies linearly with time), as shown in Fig. 19.

4. Two-Bit Quantizer with Frequency Offset Interference

Consider, now, an interference that is frequency offset from the center frequency of the PN signal. Assume that the frequency offset is large compared to the bit rate so that quantizer–correlator performance is determined by the time average of the interference in the in-phase channel ($K \cos \omega t$). Figure 20 shows the output degradation for values of $K = 10, 20, 40, 80$ relative to the signal amplitude of

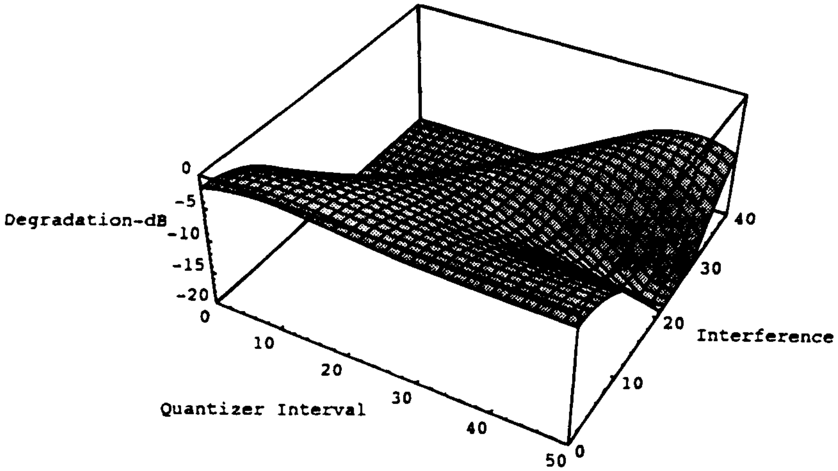


Fig. 18 Quantizer–correlator SNR degradation for the 2-bit quantizer vs both interference level and quantizer interval. The rms noise input is $\sigma = 10$.

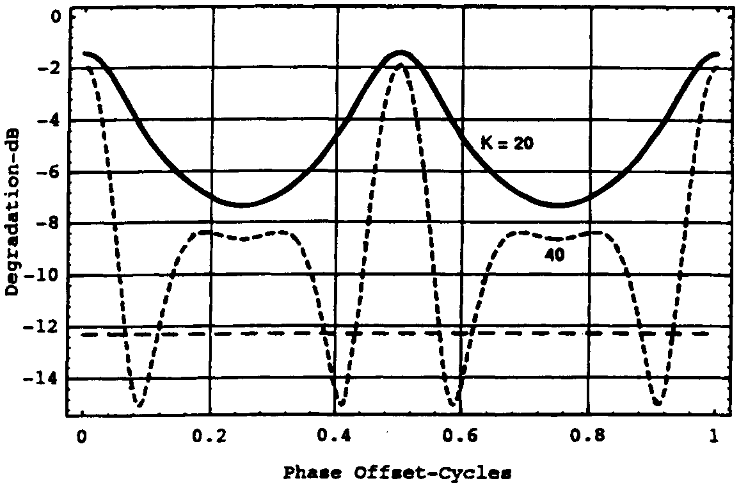


Fig. 19 Quantizer–correlator degradation for a setting of $\Delta = \sqrt{10^2 + K^2}$, with interference levels of $K = 20$ and 40. The horizontal dashed curve corresponds to the ideal quantizer degradation for $K = 40$ and a fixed phase offset $\theta = \pi/4$. The average degradation for this example with the linear quantizer would be 9.6 dB.

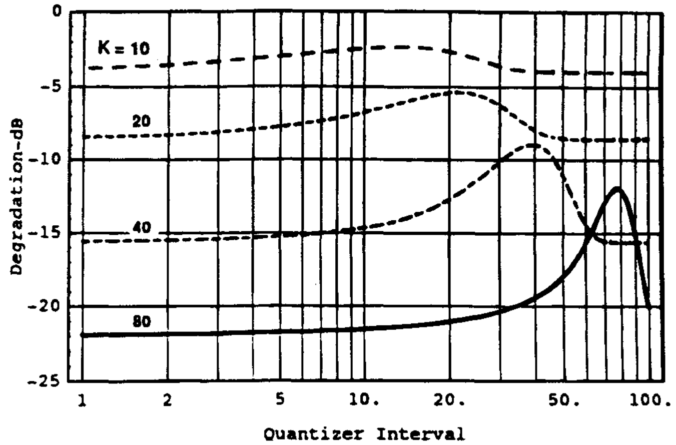


Fig. 20 Two-bit quantizer–correlator degradation in dB for four different values of the interference amplitude with frequency offset interference of the form $K \cos \omega t$. The results are given for $K = 10$, large dashes; $K = 20$, small dashes; $K = 40$, alternate dashes; and $K = 80$, solid curve.

unity and the rms noise $\sigma = 10$. The degradation in the quantizer–correlator output SNR relative to the quantizer input signal to thermal noise ratio varies with quantizer interval, as shown. Note that the optimum quantizer interval, again, has a value of $\Delta \cong \sqrt{\sigma^2 + K^2}$ and corresponds to roughly equal probable magnitude levels in the quantizer output during the interference peaks. Note, further, that the sinusoidal interference causes the quantizer output to have line components in its output. However, these components are spread by the cross-correlation operation with the C/A PN signal.

Figure 21 shows the quantizer–correlator output SNR degradation for the optimum quantizer spacing, $\Delta \cong \sqrt{\sigma^2 + K^2}$ with varying interference levels. Also shown in Fig. 21 is the degradation in quantizer output SNR relative to the quantizer input signal to thermal noise ratio that would be caused by the addition of the interference for a perfectly linear quantizer. Note that the output SNR for the quantizer correlator is improved over that of a perfectly linear quantizer–correlator for constant envelope interference above $K = 30$. At an interference level $K = 100$, the improvement is approximately 7 dB, as shown by the difference between the solid and large dashed curve. Note that the performance with a fixed amplitude interference of Fig. 17 has nearly reached an asymptote at $K = 100$; whereas, with a frequency offset, the performance degradation is still increasing in Fig. 21. Clearly, the optimum quantizer (adaptive AGC) has progressively greater performance compared to the linear quantizer as interference increases.

5. Performance with Three-Bit Quantizing—Quantizer–Correlator Degradation

Figure 22 shows the quantizer–correlator degradation for a 3-bit quantizer with a signal amplitude of $p = \pm 1$ and pure Gaussian noise with an rms value

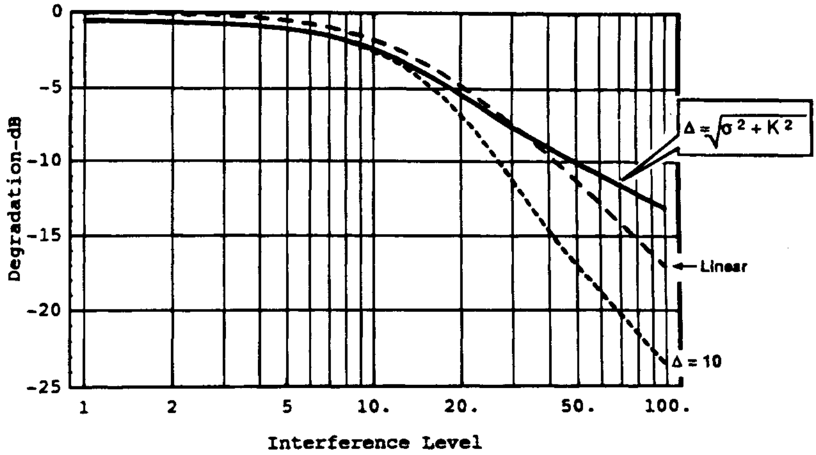


Fig. 21 Two-bit quantizer–correlator output degradation for frequency offset interference I = K cos ωt for various quantizer intervals. The solid curve represents near optimum performance with adaptive quantizer spacing $\Delta \cong \sqrt{\sigma^2 + K^2}$; the small dashed curve shows the performance with $\Delta = 10$ fixed at the optimum for no interference. The large dashed curve shows the degradation resulting from a perfectly linear quantizer and is of course equal, to the quantizer input signal to (noise + interference) power ratio.

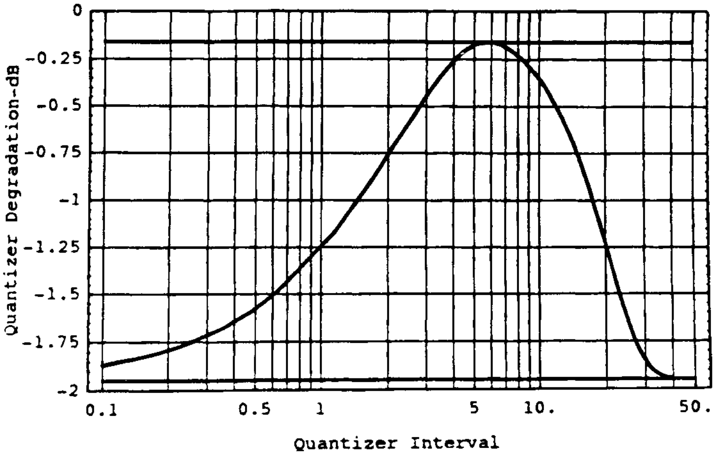


Fig. 22 Three-bit quantizer–correlator degradation in SNR for a signal level $p = \pm 1$, rms noise with $\sigma = 10$ and a quantizer interval Δ. The minimum occurs at $\Delta = 10/\sqrt{3} \approx 5.77$ and is ≤ 0.161 dB degradation. At $\Delta = 0$ or $\propto$ the degradation increases to -1.96 dB or $2/\pi$, the 1-bit quantizer result.

of $\sigma = 10$. There is no sinusoidal interference in this example. The minimum degradation occurs with a quantizer interval of $\Delta = \sigma/\sqrt{3}$ compared to $\Delta = \sigma$ for the 2-bit quantizer. The minimum degradation is 0.1613 dB vs 0.5415 dB for the 2-bit quantizer. Furthermore, for 3-bit quantization, the quantizing interval for less than 1 dB of degradation can range from $\Delta = 1.5$ to 18 or a 22 dB range. Thus, the AGC setting is considerably less sensitive than for the 2-bit quantizer, which has only a 13 dB range.

Figure 23 shows the cumulative normal probability distribution at the various quantizing intervals in terms of the magnitude probabilities. The probability of being in the magnitude interval 0 to Δ is 0.4363; Δ to 2Δ is 0.3155; 2Δ to $-\infty$ is 0.2482. Thus, the ratio of the probability of being in region 0 to Δ over that of being in Δ to 2Δ is 1.38, a little over unity. The quantizer AGC level, thus, can be set for both the 2-bit and 3-bit quantizers by properly balancing the counts in each magnitude bin (magnitude bits).

The additional quantizer–correlator output SNR degradation, beyond that of noise alone, caused by coherent interference of constant amplitude K is shown in Fig. 24 for $\sigma = 10$. The result for an ideal linear circuit is shown for comparsion. The flattening or double hump behavior appears because of the multiple quantizer magnitude levels in the 3-bit quantizer. Figure 25 gives the excess degradation for the four quantizing intervals $\Delta = \sqrt{\sigma^2 + K^2}/\sqrt{3}$, $\sqrt{\sigma^2 + K^2}$, a constant $\Delta = 10$, and the 1-bit quantizer $\Delta = 0$. For reference, the ratio of the degradation in the quantizer input SNR caused by interference; namely, $10/\sqrt{100 + K^2}$ is also shown. Again, the degradation for $\Delta = \sqrt{\sigma^2 + K^2} \approx K$ for $K >> 10$ is a special case for which the quantizing interval is almost squarely centered on the peak probability density of the interference plus noise amplitude. The degradation flattens to an asymptote at -9 dB for large K. This quantizer setting is not quite

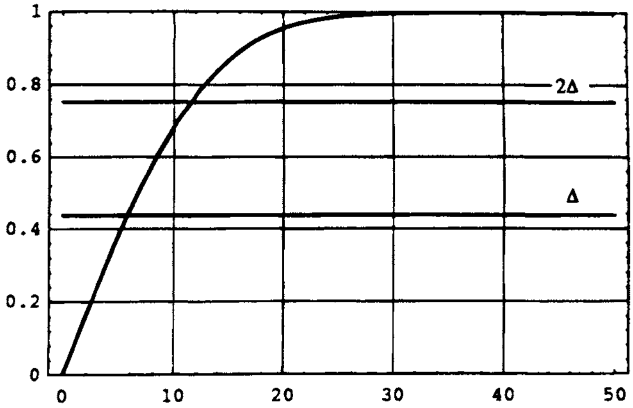


Fig. 23 Cumulative probability at the optimum quantizing interval $\Delta = 10/\sqrt{3}$, and 2Δ. Note that the ratio of the probability of being in the range 0–Δ is 0.44; whereas, the probability of being in the range Δ–2Δ is approximately 0.32. Thus, the ratio of the two probabilities can be used to set AGC level.

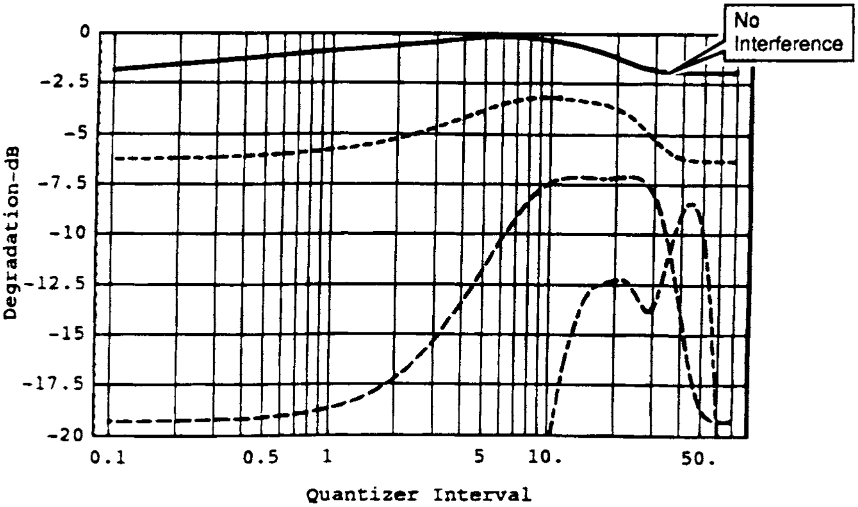


Fig. 24 Quantizer–correlator degradation for a 3-bit quantizer caused by fixed interference of amplitude $K = 0$, solid curve; $K = 10$, small dashes; $K = 20$, large dashes; and $K = 40$, alternate size dashes. The signal is $p = \pm 1$, and the noise has rms value $\sigma = 10$.

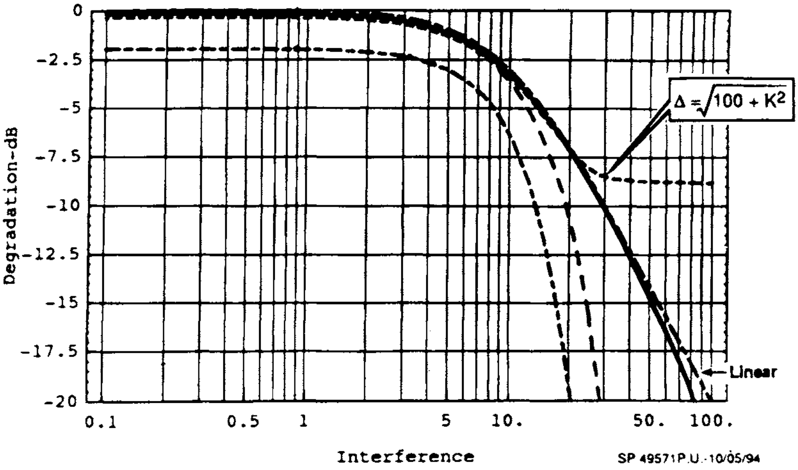


Fig. 25 Quantizer–correlator degradation for the 3-bit quantizer vs interference level for various quantizer intervals, $\Delta = \sqrt{100 + K^2} / \sqrt{3}$, solid curve; $\Delta = \sqrt{100 + K^2}$, small dash curve that levels off; $\Delta = 10$, large dashed curve. The small alternate dashed curve, the curve with the greatest degradation is the 1-bit quantizer $\Delta = 0$ for reference. A reference curve, $10/\sqrt{100 + K^2}$, the linear result, is also shown for reference with large alternate dashes.

as good for noise alone, as shown in Fig. 22. The degradation is about 0.35 dB for $\sigma = 10$, as compared to a minimum of 0.16 dB. The probability of being in the 0–Δ range increases with no interference to 0.6827, but the ratio of being in 0–Δ vs Δ–2Δ is approximately unity for high levels of interference. However, the 3-bit quantizer degradation caused by interference with amplitude $K = 10$ is nearly 11 dB less than that for a linear circuit.

Figure 26 shows the three-dimensional plot of degradation of the quantizer–correlator vs both quantizer level and interference amplitude. Again, there is a ridge of optimum quantizer level corresponding to $\Delta \cong \sqrt{K^2 + \sigma^2}$. However, with the 3-bit quantizer, there is a second ridge corresponding to the next quantizing level, and the 3-bit quantizer is not as sensitive to quantizer interval setting as is the two bit quantizer.

Figure 27 shows probability density superimposed on the quantizer intervals for $\Delta = \sqrt{100 + K^2}/\sqrt{3}$, with $K = 20$, i.e., $\Delta = 12.9$, the optimum quantizer setting for noise alone. The recommended quantizer interval for interference plus noise would be $\Delta = 22.36$.

If the coherent interference phase is varied slowly compared to the data bit rate; namely, $I = K \sin \theta$, then the excess degradation for the 3-bit quantizer is as shown in Fig. 28 as a function of the phase. The degradation is given for $K = 20$ and 40. For reference, the peak degradation for the infinite level quantizer is shown as the horizontal dashed curve, and the average degradation is 6 dB above that.

6. *Three-Bit Quantizer with Frequency Offset Interference*

Consider next a frequency offset interference $K \cos \omega t$, where the offset frequency is larger than the databit rate by several times so that power averaging occurs. Figure 29 shows the quantizer–correlator degradation for values of interference $K = 10$,

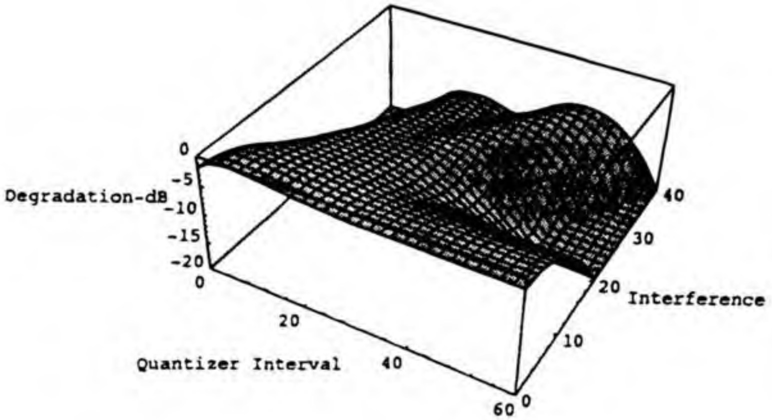


Fig. 26 Excess quantizer–correlator degradation for the 3-bit quantizer vs both quantizer interval and interference level.

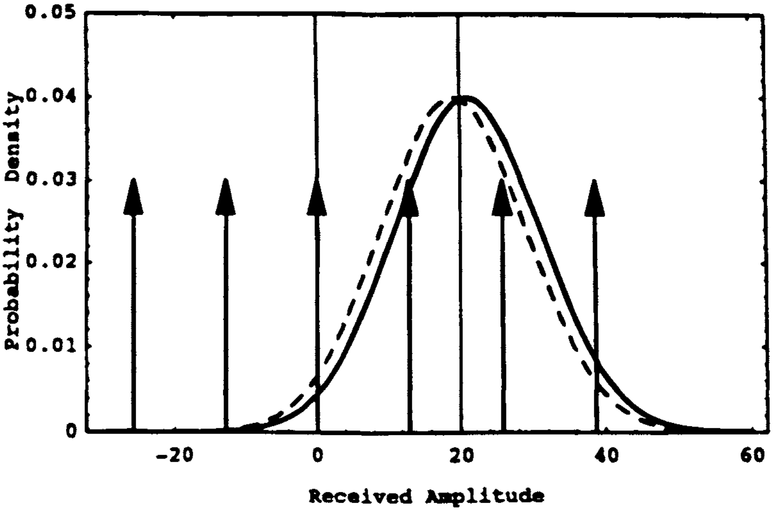


Fig. 27 Signal plus noise probability density and quantizer intervals for the 3-bit quantizer with $K = 20$ and a quantizer interval set by total power /3, i.e., $\Delta = \sqrt{100 + K^2}/\sqrt{3}$. The quantizer intervals are denoted by the arrows.

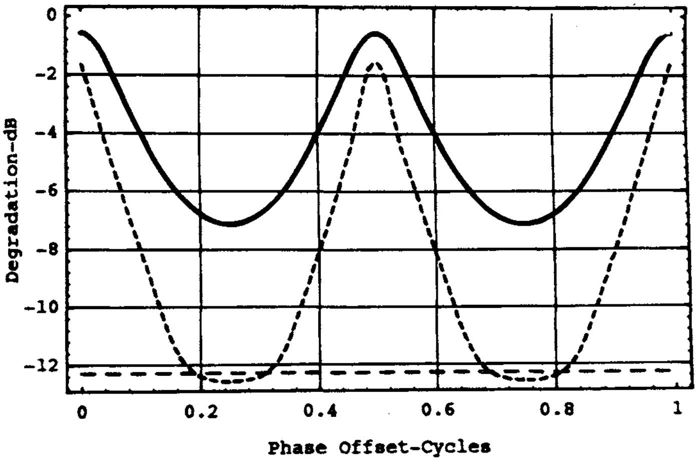


Fig. 28 Excess degradation in SNR for 3-bit quantizer for an interference $K = 20 \sin \theta$, solid curve *i* and $K = 40 \sin \theta$, small dashed curves. For reference, the peak degradation for an infinite level quantizer is shown as the horizontal dashed curve. The quantizer interval is $\sqrt{\sigma^2 + K^2}$.

AUTHOR INDEX

SUBJECT INDEX

Index Terms **Links**

Index Terms **Links**

Global Positioning System: Theory and Applications

Volume II

Edited by

Bradford W. Parkinson
Stanford University, Stanford, California
James J. Spilker Jr.
Stanford Telecom, Sunnyvale, California

Associated Editors:
Penina Axelrad
University of Colorado, Boulder, Colorado
Per Enge
Stanford University, Stanford, California

Volume 164
PROGRESS IN ASTRONAUTICS AND AERONAUTICS

Paul Zarchan, Editor-in-Chief
Charles Stark Draper Laboratory, Inc.
Cambridge, Massachusetts

Published by the
American Institute of Aeronautics and Astronautics, Inc.
370 L'Enfant Promenade, SW, Washington, DC 20024-2518

Third Printing

Data and information appearing in this book are for informational purposes only. AIAA is not responsible for any injury or damage resulting from use or reliance, nor does AIAA warrant that use or reliance will be free from privately owned rights.

ISBN 1-56347-107-8

Preface

Overview and Purpose of These Volumes

Of all the *military* developments fostered by the recent cold war, the Global Positioning System (GPS) may prove to have the greatest positive impact on everyday life. One can imagine a 21st century world covered by an augmented GPS and laced with mobile digital communications in which aircraft and other vehicles travel through "virtual tunnels," imaginary tracks through space which are continuously optimized for weather, traffic, and other conditions. Robotic vehicles perform all sorts of construction, transportation, mining, and earth moving functions working day and night with no need for rest. Low-cost personal navigators are as commonplace as hand calculators, and every cellular telephone and personnel communicator includes a GPS navigator. These are some of the potential positive impacts of GPS for the future. Our purpose in creating this book is to increase that positive impact. That is, *to accelerate the understanding of the GPS system and encourage new and innovative applications.*

The intended readers and users of the volumes include all those who seek knowledge of GPS techniques, capabilities, and limitations:

- Students attending formal or informal courses
- Practicing GPS engineers
- Applications engineers
- Managers who wish to improve their understanding of the system

Our somewhat immodest hope is that this book will become a standard reference for the understanding of the GPS system.

Each chapter is authored by an individual or group of individuals who are recognized as world-class authorities in their area of GPS. Use of many authors has led to some overlap in the subject matter which we believe is positive. This variety of viewpoints can promote understanding and contributes to our overall purpose. Books written by several authors also must contend with variations in notation. The editors of the volume have developed common notations for the important subjects of GPS theory and analysis, and attempted to extend this, where possible, to other chapters. Where there are minor inconsistencies we ask for your understanding.

Organization of the Volumes

The two volumes are intended to be complementary. Volume I concentrates on fundamentals and Volume II on applications. Volume I is divided into two parts: the first deals with the operation and theory of basic GPS, the second section with GPS performance and errors. In Part I (GPS Fundamentals), a summary of GPS history leads to later chapters which promote an initial understanding of the three GPS segments: User, Satellite, and Control. Even the best of systems has its limitations, and GPS is no exception. Part II, GPS Performance and Error Effects, is introduced with an overview of the errors, followed by chapters devoted to each of the individual error sources.

Volume II concentrates on two aspects: augmentations to GPS and detailed descriptions of applications. It consists of Parts III to VI:

- III. Differential GPS and Integrity Monitoring
- IV. Integrated Navigation Systems
- V. GPS Navigation Applications
- VI. Special Applications

Parts III and IV expand on GPS with explanations of supplements and augmentations to the system. The supplements enhance accuracy, availability, or integrity. Of special interest is differential GPS which has proven it can provide sub-meter (even centimeter) level accuracies in a dynamic environment. The last two sections (V and VI) are detailed descriptions of the major applications in current use. In the rapidly expanding world of GPS, new uses are being found all of the time. We sincerely hope that these volumes will accelerate such new discoveries.

Acknowledgments

Obviously this book is a group undertaking with many, many individuals deserving of our sincere thanks. In addition to the individual authors, we would especially like to thank Ms. Lee Gamma, Mr. Sam Pullen, and Ms. Denise Nunes. In addition, we would like to thank Mr. Gaylord Green, Dr. Nick Talbot, Dr. Gary Lennon, Ms. Penny Sorensen, Mr. Konstantin Gromov, Dr. Todd Walter, and Mr. Y. C. Chao.

Special Acknowledgment

We would like to give special acknowledgment to the members of the original GPS Joint Program Office, their supporting contractors and the original set of engineers and scientists at the Aerospace Corporation and at the Naval Research Laboratory. Without their tenacity, energy, and foresight GPS would not be.

B. W. Parkinson
J. J. Spilker Jr.
P. Axelrad
P. Enge

Table of Contents

Part IV. Integrated Navigation Systems

Part VI. Special Applications

Chapter 1

Differential GPS

Bradford W. Parkinson* and Per K. Enge†
Stanford University, Stanford, California 94305

I. Introduction

DIFFERENTIAL GPS (DGPS) is a technique that significantly improves both the accuracy and the integrity of the Global Positioning System. The most common version of DGPS is diagrammed in Fig. 1. As shown, DGPS requires high-quality GPS "reference receivers" at known, surveyed locations. The reference station estimates the slowly varying error components of each satellite range measurement and forms a correction for each GPS satellite in view. This correction is broadcast to all DGPS users on a convenient communications link. Typical ranges for a local area differential GPS (LADGPS) station are up to 150 km. Within this operating range, the differential correction greatly improves accuracy for all users, regardless of whether selective availability (SA) is activated or is not (see Chapter 11, the companion volume, on error analysis). This improvement arises because the largest GPS errors vary slowly with time and are strongly correlated over distance. Differential DGPS also significantly improves the "integrity," or truthfulness, of GPS for all classes of users, because it reduces the probability that a GPS user would suffer from an unacceptable position error attributable to an undetected system fault. (Integrity is the probability that the displayed position is within the specified or expected error boundaries.)

A. Standard Positioning Service Users

The most dramatic DGPS improvement is found for the Standard Positioning Service (SPS) when SA is activated. Although an SPS receiver itself is capable of range measurement precision of approximately 0.5 m, the normal ranging errors include slowly varying biases attributable to all six of the error classes

Copyright © 1995 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Released to AIAA to publish in all forms.

*Professor of Aeronautics and Astronautics and of the Hansen Experimental Physics Laboratory. Director of the GPS Program.

†Professor of Aeronautics and Astronautics.

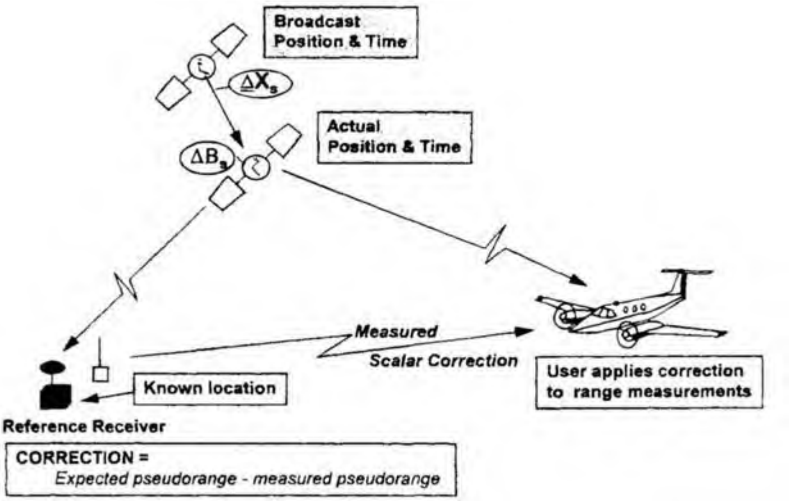

Broadcast Position & Time
ΔXs
Actual Position & Time
ΔBs
Measured
Scalar Correction
User applies correction to range measurements
Known location
Reference Receiver
CORRECTION = Expected pseudorange - measured pseudorange

Fig. 1 Local area differential GPS. The reference receiver calibrates the correlated errors in ranging to the satellites. These filtered errors are then transmitted to the user as range corrections.

described in Chapter 11 of the companion volume. These are dominated by SA, with one sigma *ranging* errors typically measured to be about 21 m. Without differential corrections, these SA-dominated biases limit the *horizontal positioning* accuracy of the SPS to 100 m (approximate 95 percentile level). With differential corrections, the SPS navigation accuracy can be improved to better than 1 m (1 σ), provided that the correction age is less than 10 s, and the user is within 50 km of the reference station. As the corrections age, or the geographic separation from the reference station increases, the accuracy of DGPS degrades. This degradation with range is graceful; thus LADGPS provides adequate accuracy for some applications at ranges of up to 1000 km.

B. Precise Positioning Service Users

As mentioned, DGPS also improves the performance of the Precise Positioning Service (PPS). Without differential corrections, the PPS is significantly more accurate than the SPS, because PPS users do not suffer from SA. In addition, PPS receivers can use measurements at *both* GPS frequencies to reduce the effect of ionospheric delays. Nonetheless, differential corrections can still provide significant improvements to the PPS accuracy, which is nominally 15 meters SEP (spherical error probable, which is the radius of a sphere that is expected to contain 50% of the errors). Expected accuracies with DGPS are about the same as SPS: they range from 1 to 5 m, depending upon the system design.

C. Major Categories of Differential GPS

There are many DGPS techniques and applications.[1,2] The major techniques are broadly characterized in the following subsections.

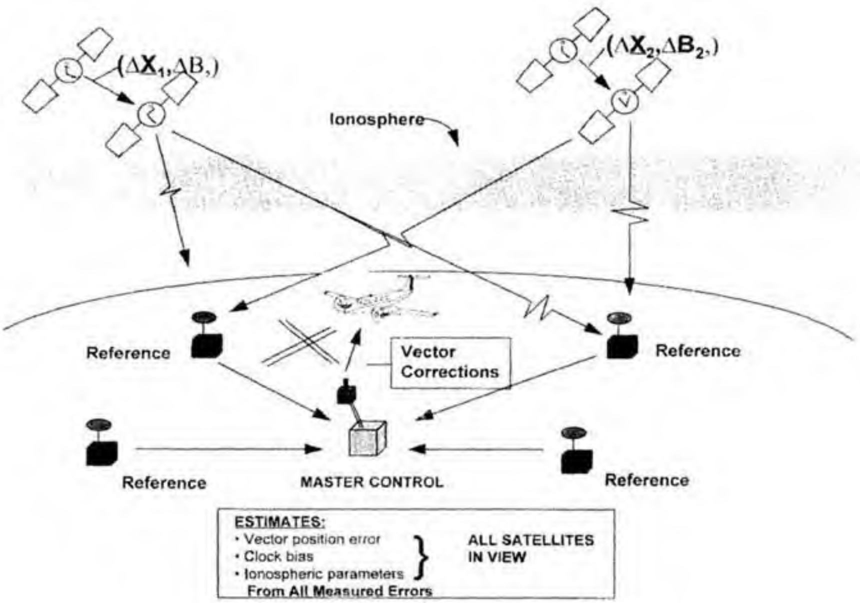


Fig. 2 Wide area differential GPS concept. One type, sponsored by the FAA, is known as the wide-area augmentation system (WAAS).

1. Local Area Differential GPS

Most DGPS systems use a single reference station to develop a *scalar* correction to the code-phase* measurement for each satellite. This approach is shown in Fig. 1. If the correction is delivered within 10 s, and the user is within 1000 km, the user accuracy should be between 1 and 10 m. This capability (shown in Fig. 3) is detailed further in Sec. II of this chapter. An additional technique uses inexpensive, ground-based transmitters that broadcast a GPS signal at the L_1 or L_2 frequency. These are called pseudosatellites or *pseudolites* (PL) and act as an additional ranging source as well as a datalink. Pseudolites provide significant improvements in geometry[3] and accuracy; one technique is described under test results and discussed in a later chapter on precision landing.

2. Wide Area Differential GPS

As shown in Fig. 2, networks of reference stations can be used to form a *vector* correction for each satellite. This vector consists of individual corrections for the satellite clock, three components of satellite positioning error (or ephemeris), and parameters of an ionospheric delay model. The validity of this correction

*The modern technique for receiver code-phase measurements is to use "carrier aiding," which filters the noisy code-phase measurements with the smoother carrier measurements. This is not to be confused with pure carrier-tracking techniques described further later.

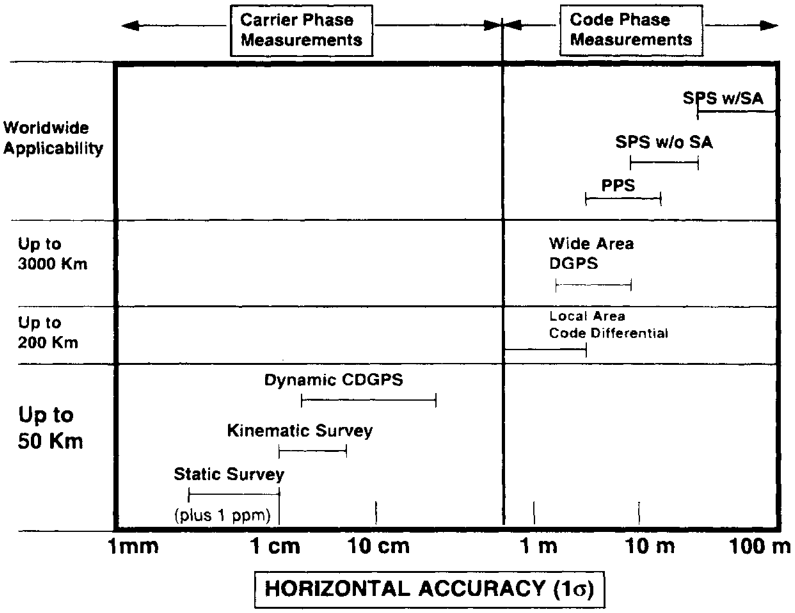


Fig. 3 Summary of expected differential GPS concepts and accuracies.

still decreases with increased *latency** or age of the correction. However, compared to a scalar correction, a vector correction is valid over much greater geographical areas. This concept is called wide area DGPS, or WADGPS.[4] Such networks will be used for continental or even world-hemisphere coverage, because they require many fewer reference stations than a collection of independent systems with one reference station each. Moreover, they require less communication capacity than the equivalent network of LADGPS systems. Wide area GPS is a subject unto itself, and it is described in detail in Chapters 3 and 4 of this volume.

3. *Carrier-Phase Differential GPS*

Users with very stringent accuracy requirements may be able to use a technique called carrier-phase DGPS or CDGPS. These users measure the phase of the GPS carrier relative to the carrier phase at a reference site; thus achieving range measurement precisions that are a few percent of the carrier wavelength (typically about one centimeter). These GPS phase comparisons are used for vehicle attitude determination and also in survey applications, where the antennas are separated by tens of kilometers. If the antennas are fixed, then the survey is called *static*, and millimeter accuracies are possible, because long averaging times can be used to combat random noise. If the antennas are moving, then the survey is *kinematic*,

*Latency is the total time from the reference station measurement of error to the actual application in the user receiver. It includes the calculation time and any communications delay.

and shorter time constants must be used with some degradation of accuracy. These static and kinematic capabilities are included in Fig. 3. Several carrier-phase techniques for aircraft precision landing have also been demonstrated. Carrier-phase DGPS is introduced in Sec. IV of this chapter and is further described in Chapters 4, 15, 18, and 19 of this volume.

4. *Organization of the Chapter*

This chapter introduces DGPS, and many of the remaining chapters apply or further develop this important technique. Section I of this chapter describes the measurements of a code-phase differential system. In Sec. III the error analysis for a LADGPS is developed. Accuracy degradation for "aged" corrections and for user displacements from the reference station are quantified. Section IV introduces CDGPS by describing GPS phase interferometry for attitude determination as well as static and kinematic survey. It also introduces techniques for resolving the λ, or wavelength ambiguity, which must be determined to realize centimeter-level accuracies. Section V describes standardized data formats for the transmission of local area differential corrections, and Sec. VI provides an overview of DGPS broadcast systems. Section VII provides a small sample of the huge number of DGPS field results reported in the literature.

II. Code-Phase Differential GPS

It is useful to summarize the expected user errors in a form that allows analysis of differential system accuracy. Errors can be categorized as either correlated between receivers or uncorrelated. Only the correlated errors can be corrected with DGPS. Even the nominally correlated errors lose that correlation if they are significantly delayed in application (*temporally decorrelated*) or are applied to a receiver significantly separated from the reference station (*geographically decorrelated*). This section provides estimates of these decorrelation factors.

A. User Errors Without Differential GPS

This section draws heavily on the development of Chapter 11, the companion volume, which should be used as a reference. A code-tracking receiver actually measures the raw difference between the user's biased clock and the transmitted time of the start of the satellite code phase (which is part of the satellite message). This quantity is called raw *pseudorange* ρ. With the speed of light c used to convert time to distance, this is expressed as follows:

$$\rho = c \cdot (t_{\mathrm{Au}} - t_{\mathrm{Ts}}) \qquad (1)$$

where t_{Au} = *A*rrival time measured by the *u*ser, s; t_{Ts} = uncorrected value of *s*atellite *T*ransmission time, s; and u,s represents the user and the sth satellite. This is measured (or raw) pseudorange, which equals the true range D from the user u to satellite s plus an unknown offset between the user clock b_u and the satellite clock B_s. Additional time delays are caused by the ionosphere I and the troposphere T, as well as noise, multipath, and/or interchannel errors in the user's receiver v:

$$\rho = D + c \cdot (b_u - B) + c \cdot (T + I + v) \tag{2}$$

The true geometric range (in *xyz* Cartesian coordinates) is given by the following:

$$D = \sqrt{(x_s - x_u)^2 + (y_s - y_u)^2 + (z_s - z_u)^2}$$

This can also be written in a more convenient way for calculations:

$$D = |\bar{r}_s - \bar{r}_u| = \bar{1}_s \cdot [\bar{r}_s - \bar{r}_u] \tag{3}$$

where $\bar{r}_s$ = true satellite position (included in user message); $\bar{r}_u$ = true user position; and $\bar{1}_s$ = true unit vector from users to satellite.

As an aside, Eq. (2) is modified as follows if carrier phase is the basic measurement. The carrier cycles are counted and converted to range so that the quantity $\phi_{u,s}$, formed by counting zero crossings of the reconstructed radio frequency (rf) carrier, is in meters.

$$\phi_{u,s} = D + c \cdot (b_u - B_s) + c \cdot (T - I + v_{u,s}^{(\phi)}) + N_{u,s} \cdot \lambda \tag{4}$$

Note that the *sign* of the ionospheric *group* delay is changed for this *phase* delay. Also note the addition of $N_{u,s} \cdot \lambda$, where N is an unknown integer that counts carrier wavelengths, and λ is the carrier wavelength of 19.2 cm. Tropospheric errors affect both types of measurements equally. Also note the superscript on the receiver noise. Carrier measurement noise is not the same as that measured for the code; in fact, it is usually orders of magnitude smaller.

Returning to code-phase measurements, the measurement represented by Eq. (2) is adjusted to form *corrected pseudorange.* This is formed by correcting the measurement for estimates of some of the raw errors:

$$\rho_{c_{u,s}} = \rho_{u,s} + c \cdot \hat{B}_s - c \cdot (\hat{T}_s + \hat{I}_s) \tag{5}$$

where the ($\wedge$) is used to denote estimates of satellite timing error or estimates of ionospheric or tropospheric delays.

A user without DGPS, then, forms four or more of these measurements into a matrix equation as developed in Chapter 11 of the companion volume:

$$\boldsymbol{G}\hat{\bar{x}} = (\mathbf{A} \cdot \bar{R} - \bar{\rho}_c) \tag{6}$$

and solves* for estimated position as follows:

$$\hat{\bar{x}} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T(\boldsymbol{A} \cdot \bar{R} - \bar{\rho}_c) \tag{7}$$

*Note that this pseudoinverse collapses to $\hat{\bar{x}} = \boldsymbol{G}^{-1}(\boldsymbol{A} \cdot \bar{R} - \bar{\rho}_c)$ if the number of satellites equals four.

where

$$\hat{\bar{x}}_{4\times 1} \equiv \begin{bmatrix} \hat{\bar{r}}_u \\ -c\cdot\hat{b}_u \end{bmatrix}; \qquad \boldsymbol{G}_{n\times 4} \equiv \begin{bmatrix} \hat{\bar{1}}_{s1}^T & 1 \\ \hat{\bar{1}}_{s2}^T & 1 \\ \vdots & \vdots \\ \hat{\bar{1}}_{sn}^T & 1 \end{bmatrix};$$

$$\boldsymbol{A}_{n\times 3n} \equiv \begin{bmatrix} \hat{\bar{1}}_{s1}^T & & & 0 \\ & \hat{\bar{1}}_{s2}^T & & \\ & & \ddots & \\ 0 & & & \hat{\bar{1}}_{sn}^T \end{bmatrix}; \qquad \bar{\boldsymbol{R}}_{3n\times 1} \equiv \begin{bmatrix} \hat{\bar{r}}_{s1} \\ \hat{\bar{r}}_{s2} \\ \vdots \\ \hat{\bar{r}}_{sn} \end{bmatrix}$$

Note that $\boldsymbol{G}$, the geometry matrix, is determined by the estimated directions to each of the visible satellites. $\boldsymbol{A}$ is a matrix of the satellite locations which have been received as part of the satellite broadcast, and $\hat{\bar{\rho}}_c$ is the corrected pseudorange to each satellite, arranged as a vector. We can derive the following vector to be the pseudorange error ($\Delta\bar{\rho}$):

$$\Delta\bar{\rho} \overset{\Delta}{=} c\cdot(-\Delta\bar{B} + \Delta\bar{I} + \Delta\bar{T} + \bar{\boldsymbol{\nu}}) + \epsilon\cdot(\bar{R} - \bar{P}) + \boldsymbol{A}\cdot\Delta\bar{R} \tag{8}$$

where the first four Δ's are the residual errors caused by satellite clock (including SA), ionosphere, and troposphere (after any corrections). The vector $\boldsymbol{\nu}$ includes receiver errors and multipath, and we also define

$$\epsilon_{n\times 3n} \equiv \begin{bmatrix} \Delta\hat{\bar{1}}_{s1}^T & & & 0 \\ & \Delta\hat{\bar{1}}_{s2}^T & & \\ & & \ddots & \\ 0 & & & \Delta\hat{\bar{1}}_{sn}^T \end{bmatrix}; \qquad \bar{P}_{3n\times 1} \equiv \begin{bmatrix} \bar{r}_u \\ \bar{r}_u \\ \vdots \\ \bar{r}_u \end{bmatrix}$$

Note that in Eq. (8), the satellite position error is the last term, and the error in calculation caused by an erroneous unit vector is the next-to-last term. This is the error in the range to the satellite. The error in the user position calculation is then given as follows:

$$\Delta\bar{x} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\Delta\bar{\rho} \tag{9}$$

The major purpose of all DGPS systems is to estimate the user's stand-alone ranging error $\Delta\hat{\bar{\rho}}_c$ so that a more accurate pseudorange can be used to estimate position.*

*There have been differential systems that corrected position rather than pseudorange. This is not a good design approach, because it assumes that both reference and user employ the same set of satellites, or else that all combinations of satellites are provided as part of the correction message.

Table 1 GPS errors with selective availability for the Standard Positioning Service

Error Source	1 sigma position error (m)		
	BIAS	RANDOM	TOTAL
Ephemeris Data	2.1	0.0	2.1
Satellite Clock	20.0	0.7	20.0
Ionosphere	4.0	0.5	4.0
Troposphere	0.5	0.5	0.7
Multipath	1.0	1.0	1.4
Receiver Measurement	0.5	0.2	0.5
Reference Station Errors	0.0	0.0	0.0
Pseudo-Range Error (RMS)	20.5	1.4	20.6
Filtered PRE (RMS)	20.5	0.4	20.5
Total Vertical Error	VDOP = 2.5		51.4 m
Total Horizontal Error	HDOP = 2.0		41.1 m

Table 1 summarizes the errors that contribute to both stand-alone GPS and DGPS before the DGPS correction. As shown, the total pseudorange error can be approximated by taking the square root of the sum of the individual errors squared. The total pseudorange error (1σ) for a GPS user without differential corrections is approximately 21 m. Clearly, SA is the dominant error source for the user without differential corrections. This table should be compared to Table 8 of this chapter, which shows the residual errors after applying DGPS corrections.

B. Reference Station Calculation of Corrections

The reference receiver turns the problem around. Its receiver antenna is located in a known position* relative to the desired reference frame. It then solves Eq. (6) for the unknown corrections to the raw pseudorange vector $\bar{\boldsymbol{\rho}}$ using†: $\mathbf{G}\hat{\bar{\mathrm{x}}}_T = (\mathbf{A} \cdot \bar{R} - \bar{\boldsymbol{\rho}} - \Delta\hat{\bar{\rho}}_R)$, thus we have the following:

$$\Delta\hat{\bar{\rho}}_R = \mathbf{A} \cdot \bar{R} - \bar{\boldsymbol{\rho}} - \mathbf{G}\hat{\bar{x}}_T$$

This is the fundamental reference station calculation. This $\Delta\hat{\bar{\rho}}_R$ (estimated ranging error) is then transmitted to the user, who applies it to his measured psuedoranges.

*The choice of reference frames is up to the system designer. Any convenient frame is acceptable, but all reference points should be located to the same desired accuracy. For example, location of a runway relative to the reference antenna should have a consistent level of accuracy.

†It is best for both the reference and the user to use raw pseudoranges. Any corrections applied must be consistent at both places.

The "known" reference station position is *four-dimensional.* That is, it includes a local time correction. Any consistent timing error for *all* pseudorange corrections will only affect the user clock (see Chapter 11, the companion volume, for the reason). Assuming that the user is interested only in the user's geographical position, this clock correction is, therefore, arbitrary. This time bias is usually slowly "steered" so that the magnitude of the largest correction is minimized.

C. Application of Reference Correction

Several important points must be made about applying this correction:

1) Any pseudorange corrections that are in addition to the reference correction must be agreed upon by both user and reference, and they must be applied in exactly the same way. For example, in the formation of the "raw" pseudorange measurements $\bar{\rho}$, the broadcast ionospheric delay correction must be the same for both the user and the reference station. The safest course of action is for neither to apply corrections. Assuming this is the case, the user applies the correction that has been received ($\Delta\hat{\bar{\rho}}_R$) to the measured raw psuedoranges. The user's fundamental equation becomes $\boldsymbol{G}\hat{\bar{x}}_u = [\boldsymbol{A} \cdot \bar{R} - (\bar{\rho} + \Delta\hat{\bar{\rho}}_r)]$. The solution for the $n = 4$ case is $\hat{\bar{x}}_u(t) = \boldsymbol{G}^{-1} [\boldsymbol{A} \cdot \bar{R} - (\bar{\rho}\ (t) + \Delta\hat{\bar{\rho}}_R(t - \tau)]$. The τ has been introduced to highlight the delay between measurement and application.

2) The reference and the user *must be using the same satellite ephemeris.* Because these are periodically revised, a well-designed DGPS system will continue to broadcast corrections based on both old and new ephemerides during transition periods. The user can then use either correction while he completes gathering the new ephemeris data.

3) The reference station must take great care to not introduce additional user errors by including effects that are not measured by the user. An example would be multipath error induced by reflections from buildings near the reference station antenna.

4) The time of the reference station correction should be passed to the user as part of the correction message. This will allow the user to both evaluate integrity and properly apply any time-rate-of-change information.

III. Analysis of Differential GPS Errors

In general, the reference station gathers corrections that are geographically separated from the user and are delayed in application. A first-order relationship between the reference correction and the user's best correction is as follows:

$$\Delta\bar{\rho}_u \cong \Delta\hat{\bar{\rho}}_R + \frac{\partial(\Delta\bar{\rho})}{\partial\bar{x}} \cdot \delta\bar{x} + \Delta\dot{\bar{\rho}}_R \cdot \Delta t + \delta\bar{\rho}$$

The first term $\Delta\hat{\bar{\rho}}_R$ is the correction estimated by the reference station. The next three terms represent deviations from a perfect correction. These errors are referred to as types 1, 2, and 3, and they are defined as follows:

- ***Type 1: Decorrelation with Distance.*** The term $[\partial(\Delta\bar{\rho})]/(\partial\bar{x}) \cdot \delta\bar{x}$ is the error attributable to the vector gradient of corrections (the decorrelation with distance from the reference site).

- ***Type 2: Decorrelation with Time.*** The term $\Delta\dot{\rho}_R \cdot \Delta t$ is the error attributable to the time rate of change of the corrections (decorrelation with time). This effect is frequently called *latency.* To cope with time decorrelations, most DGPS systems broadcast the measured time rate of change of corrections as part of the communicated message. This first-order correction usually achieves an accuracy of about 0.5 m for a 10 s delay. Higher-order derivatives can be transmitted, but their estimates are noisy, and the prediction process is deliberately made difficult by the high-frequency changes induced by SA.

- ***Type 3: Uncorrelated Errors (not correctable with DGPS).*** The last term represents errors at the user that are not correlated with those measured at the reference. This term can be viewed as the error for user and reference if they were next to each other and there were no delay in application of the corrections. Note that type 3 errors at *both* the user and reference station contribute to the total DGPS user error.

These error types are not necessarily mutually exclusive. For example, ionospheric errors are both type 1 and type 2, because they decorrelate with both time and distance.

The subsequent sections analyze each of these error sources and present estimates of residual errors after differential corrections. For each, the distance and time decorrelation factors are also estimated. The preceding definitions of decorrelation types are used in these discussions.

A. Receiver Noise, Interference, and Multipath Errors for Differential GPS

Receiver noise, interference, and multipath are grouped together because *they constitute the noise floor* for DGPS. These errors are almost totally of type 3. They have very short decorrelation distances; thus noise, interference, and multipath at the reference station are not usually correlated with those effects at the mobile receiver.

Special care must be taken with type 3 errors in the reference station. Any effects in the reference correction *will be directly added to the user error, because they will be uncorrelated errors that are included in the broadcast correction.* Therefore, the elimination of these effects in the reference receiver is a primary design goal. Fortunately, two techniques—carrier aiding and narrow correlator spacing—can minimize these effects. Their use has significantly reduced this DGPS noise floor. Both techniques can be used with mobile receivers as well. These are discussed in the following two sections.

Multipath arises when GPS signals travel over multiple paths from the satellite to the receiver. Some of the signals are delayed relative to the "direct" signal, because they have traveled paths that include a reflection. The reflecting object might be a building, ship, aircraft, or truck, or it might be the surface of the sea or of a runway. In general, the strongest reflections occur close to the receiver. If these reflected signals are delayed by more than 1.5 μs (about 500 m of increased path length), they will be suppressed in the decorrelation process, because the autocorrelation of the C/A-code is nearly zero for delays greater than $1\frac{1}{2}$ chips. However, if they are delayed by less than 1.5 μs, their impact depends upon their amplitude, the amount of delay, and the persistence of the reflection. This persistence can be quantified as the *correlation time.*

Multipath errors, particularly in the reference station, should be the major issue in selecting and siting reference antennas. Certain modern antennas have substantial improvements in sidelobe suppression, which helps further eliminate multipath before it can enter the receiver. Avoiding antenna sites close to reflective materials can also help greatly. These considerations should be regarded as the primary defense against multipath.

1. Random Errors and Carrier Aiding

Code and carrier measurements both suffer from random observation noise, which is denoted v for the code phase and ϕ for the carrier. These random variables model the impact of thermal noise, multiple access interference, and multipath. In the absence of multipath, the standard deviation of the carrier noise is 1 cm or less compared to over 1 m in the unaided code. Therefore, the carrier-phase measurement is much more precise than the code measurement, but the carrier measurement does suffer from the mentioned integer ambiguity $N \cdot \lambda$ caused by the unknown number of carrier phase cycles between the user and the satellite.

Carrier aiding is a technique that uses the precision of the carrier observations to smooth the observed code-phase measurements. The following (fading memory) recursion is an example of a filter that is used:

$$\hat{\rho}_{u,s}(t_k) = \frac{1}{L}\,\rho_{u,s}(t_k) + \frac{L-1}{L}\,(\hat{\rho}_{u,s}(t_{k-l}) + \phi_{u,s}(t_k) - \phi_{u,s}(t_{k-1}))$$

where

$$\hat{\rho}_{u,s}(t_0) = \rho_{u,s}(t_0)$$

The first term of the recursion is the current code-phase measurement weighted by $1/L$, where L is a large number, perhaps 100 or 200. The current code-phase measurement receives a relatively low weighting because the carrier-phase difference in the second term predicts the future value of the pseudorange with very high accuracy. The forward prediction does not suffer from any integer ambiguity because the *carrier difference* is used. Moreover, under most conditions, well-designed GPS phase–lock-loops (PLL) rarely suffer from cycle slips that would degrade accuracy.

This carrier-aiding technique should not be confused with rate-aiding techniques, which use integrated Doppler measurements. Indeed, the carrier-phase measurements maintain phase coherency and do not suffer from accumulated error growth caused by accumulated measurement noise. Nonetheless, the forward prediction will eventually degrade because of code–carrier divergence (attributable to the ionosphere). In fact, the weighting constant L must be carefully chosen to balance the very low noise of the carrier measurements with the accumulation of code–carrier divergence. In the absence of significant multipath, carrier aiding bounds the standard deviation of the pseudorange error to a few tenths of a meter.

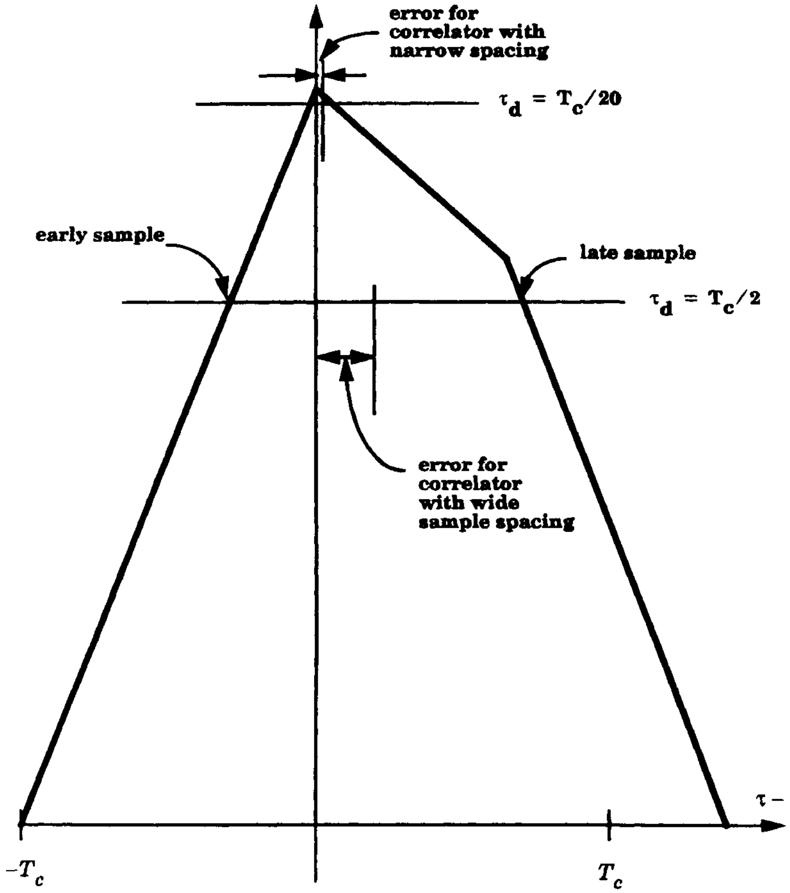


Fig. 4 Use of narrow correlator spacing to mitigate the effects of multipath.

If L_2 frequency measurements are used to track ionospheric delays, the time constant of the filter can be much greater, thus better precision can be achieved.

On a moving vehicle, the multipath correlation time may be very small (because the differential path length is changing rapidly), and carrier aiding may be quite effective in averaging any disturbances. In fact, antenna designs that intentionally randomize the phase difference between the direct and delayed signals are being considered for moving platforms. At fixed sites, the correlation time tends to be significantly longer; thus carrier aiding is not as effective. Of course, the antenna can be more carefully located and designed to have very low gain at low or negative elevation angles in order to combat multipath at a fixed location.

2. *Multipath and Narrow Correlator Spacing*

In addition to antenna selection and siting, there is a receiver-processing technique that can be used to mitigate (somewhat) the effects of multipath.

As discussed by Refs. 5–7, multipath interference can be reduced further by minimizing the time between early and late correlator samples. This is known as *narrow correlator spacing*. A sample of the correlation function in the presence of multipath is shown in Fig. 4. As shown, the multipath interference distorts the shape of the correlation function, which is symmetric in the absence of multipath. The advantage of narrow correlator spacing can be seen in the figure. If the correlators are separated by 1.0 T_c, then the early and late samples will settle at the location indicated, and the error caused by multipath can be quite large. In contrast, if the correlator spacing is $0.1T_c$, then the correlator samples will settle near the peak, and the error will generally be smaller than 1 m.

3. Summary of Receiver Noise and Multipath Errors

Table 2 summarizes this class of ranging errors under the following assumptions:

1) The user has a state-of-the-art, multichannel receiver with a modern digital signal processor.

2) The reference station has taken great care to reduce multipath susceptibility, as described in the preceding subsections.

3) The magnitude of the random error in the reference station is also found in the user's receiver (but is uncorrelated) and multiplies this statistic by the square root of 2.

Table 2 Errors in DGPS caused by receiver noise and multipath for a well-designed user equipment receiving corrections from a well-designed reference station[a]

	Without DGPS Correction with or without SA Clock Dither		Zero Baseline Zero Latency DGPS (Type 3)		Decorrelation with Latency (Type 2)		Geographic Decorrelation (m/100 km) (Type 1)
	Bias (m)	Random (m)	Bias (m)	Random (m)	Vel. (m/s)	Accel. (m/s^2)	
Receiver noise	0.5	0.2	0.5	0.3	0.0	0.0	0.0
Multipath	0.3 to 3.0	0.2 to 1.0	0.3 to 3.0	0.2 to 1.0	0.0	0.0	0.0

[a] All effects are one-sigma errors. Bias is a steady value with persistence of more than 5 s. Random is the measurement-to-measurement variation in the user or reference receiver, sometimes called "white noise." Using carrier smoothing or other averaging techniques, the random errors can be significantly reduced in modern receivers.

Because these errors are all type 3, there is no decorrelation because of latency or separation, as indicated in the three right-hand columns in the table. (Later tables follow this same format.)

B. Satellite Clock Errors for Differential GPS

Satellite clock errors are differences in the true signal transmission time and the transmission time implied by the navigation message. In the absence of SA, these errors are small and change slowly. During periods when SA was not activated, clock errors of about 1–2 m and correlation times of about 5 min have been measured.

In the presence of SA, clock errors of 20–30 m are not unusual. *Differential corrections can be very effective against clock errors, because their validity decreases only with time and not with distance.* In other words, this error is exclusively type 2. Because SA has relatively large, fairly random velocity and acceleration magnitudes, it totally dominates the latency-induced error growth. The DGPS positioning error, therefore, grows as the DGPS correction ages.

Most DGPS implementations are relatively simple and predict future values of the pseudorange correction from the current values of the pseudorange and its rate. In this case, the residual pseudorange error growth attributable to SA is approximated as $1/2\ at^2$, where a is the rms acceleration (a random variable) of SA, and t is the age of the correction in seconds. Typically, SA has exhibited an acceleration a error in range (1σ) of about 4 mm/(s)^2. Consequently, if the latency is 10 s, then the pseudorange error (1σ) attributable to SA is expected to grow to approximately 0.2 m.

Somewhat more accurate DGPS systems have used system identification techniques in real time to build more sophisticated models of SA. Three good examples are the following: 1) a second-order Gauss–Markov model[8]; 2) an autoregressive moving average (ARMA) model[9]; and 3) a technique using autoregressive (AR) models and lattice filters.[10]

All of these models are still limited by the deliberate uncertainties in the *true*, presumably nonlinear SA model. However some improvement can be realized by transmitting the particular *estimator* model's parameters to the users as well as the measured current state of the SA offset. The user's receiver can then reconstruct the current approximation to the SA model to make more sophisticated predictions.

Figure 5 shows the standard deviation of the DGPS range error as a function of the age of the correction. The error for a user without differential corrections is the horizontal line at about 34 m.* The curve marked "two state" is for a differential user who employs the simple prediction based on the current value of SA and its rate. Finally, the curve marked "optimal prediction" is for a differential user who uses a more complete estimator model (such as referenced above) for predicting future values of SA. It assumes that the estimator model parameters have been "optimally" estimated by including the known statistics of SA.

As shown in Fig. 5, DGPS can reduce the pseudorange error provided that the correction is delivered promptly. Note that the initial growth of the error

*This value is larger than typical SA errors which are closer to 23 m (1σ).

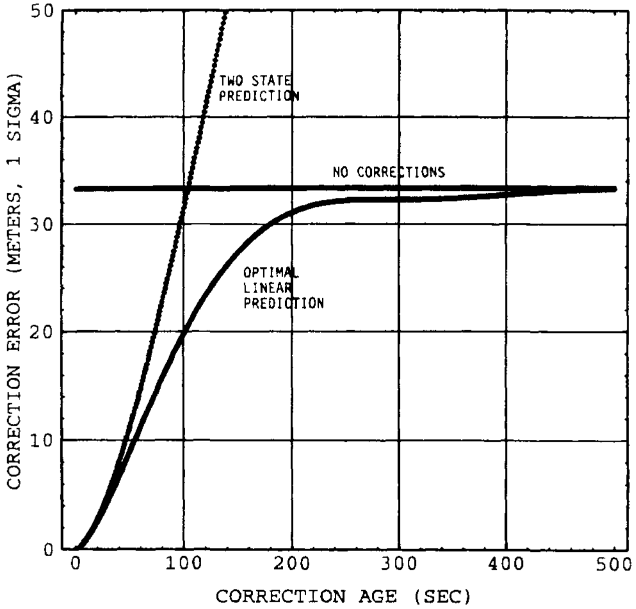


Fig. 5 Growth of horizontal differential GPS range error caused by selective availability as a function of the age of the correction.

(first 30 s) for both cases is parabolic: it grows as time squared. In fact, a delay of 20 s will lead to an error standard deviation of about 3 m in pseudorange, which corresponds to a (2 drms) positioning error of approximately 10 m.

The error for the optimal prediction never exceeds that for a user without corrections. In contrast, two-state prediction will give larger errors than nondifferential processing if the rate term is used to predict too far into the future. However, the error for the two-state prediction is very close to that for optimal prediction for smaller correction ages. For example, the differential error grows to over 10 m if the correction age exceeds 50 s, but for ages less than 50 s, the two-state prediction is almost as good as the optimal prediction.

Table 3 summarizes the statistics for DGPS satellite clock errors.

C. Satellite Ephemeris Errors for Differential GPS

As mentioned earlier, the navigation message contains errors. We have asserted that errors in the satellite clock data can be corrected by DGPS. Furthermore, these *clock corrections* are valid regardless of the distance between the monitor and the user. In other words, there is no decorrelation with displacement between reference and user. On the other hand, if the errors are in the satellite ephemeris data, then the validity of the corrections will decrease as the distance between the user and reference station increases.

In the appendix to this chapter, there is a detailed development of the *scalar* errors in DGPS ranging corrections as a vector function of the *vector* errors in

Table 3 Clock errors before and after differential GPS corrections[a]

	Without DGPS Correction with or without SA Clock Dither		Zero Baseline Zero Latency DGPS (Type 3)		Decorrelation with Latency (Type 2)		Geographic Decorrelation (m/100 km) (Type 1)
	Bias (m)	Random (m)	Bias (m)	Random (m)	Vel. (m/s)	Accel. (m/s^2)	
Satellite Clock Errors	21.0	0.1	0.0	0.14	0.21	0.004	0.0

[a] The random error is increased, because the noise in both the user and reference receivers are added after the corrections.

satellite position and the vector displacement (sometimes called the *baseline*) between reference station and user. The correct first-order expansion for these errors is the following:

$$\Delta E \triangleq \frac{-\Delta\bar{R}^T\cdot\bar{d}}{R} \tag{10}$$

where $\Delta E \triangleq$ scalar error in DGPS correction; $\bar{d} \triangleq [\Delta\bar{r}_{r,u} - (\bar{1}_s \cdot \Delta\bar{r}_{r,u})\bar{1}_s]$; $\Delta\bar{R} \triangleq$ satellite position error vector; $R \triangleq$ range from reference station to satellite; $\bar{1}_s \triangleq$ unit vector from reference station to satellite; and $\Delta\bar{r}_{r,u} \triangleq$ displacement from reference station to user.

Equation (10) shows the decorrelation of errors as the user moves away from the reference station. If we consider the magnitudes of these quantities, the error in the DGPS correction is bounded by the following:

$$\max|\Delta\hat{\rho}_R - \Delta\hat{\rho}_u| = \Delta E \leq \frac{|\Delta\bar{r}_{r,u}|\cdot|\Delta\bar{R}|}{R}$$

This equation is valid for separations of less than 1500 km (the usual case for scalar corrections); larger separations are treated below. This error may be quite conservative depending upon the exact orientation of the vectors in Eq. (10). Further feeling for this "worst-case" relationship is given by Fig. 6, which plots worst-case user position error (after DGPS correction) vs satellite position error. For example, a 100-m satellite positioning error at 100 km separation between user and reference produces user errors *in the worst case* of less than 1 m.

1. Selective Availability Effects on Ephemeris

Typical satellite ephemeris errors are usually less than 10 m. Although SA could be applied by creating errors in the ephemeris message, this technique

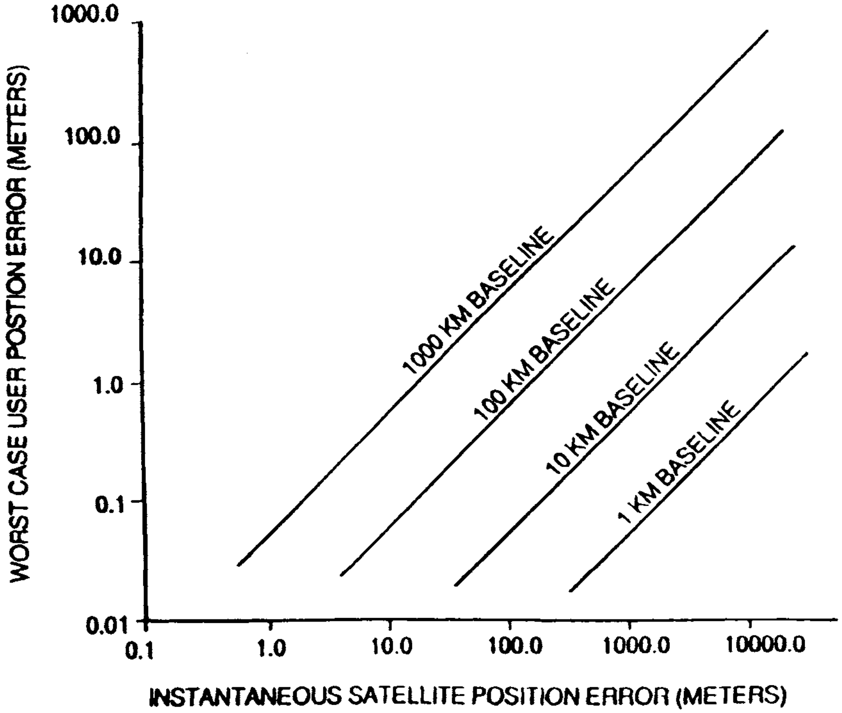


Fig. 6 Worst-case differential GPS errors for various distances from reference to user.

apparently has not been used. This is because any errors in the ephemeris would be slowly changing, and hence, strongly correlated over many minutes. Therefore, corrections for these errors would be valid for extended periods, which defeats the purpose of SA. It is assumed that the worst case, if SA were used, would limit the ephemeris message error to 100 m.

2. *Maximum Errors*

The maximum separation between a user and a reference station that can still have common view of all possible satellites is determined by their minimum elevation angles, or *mask* angles. Figure 7 shows a reference station and a user with a central angle separation of 142 deg (2.48 radians). This is the maximum common view separation, assuming the user and reference station both have elevation mask angles of 5 deg. The .maximum errors caused by this extreme separation have three components that correspond to the three components of satellite ephemeris error before differential corrections. Even at this extreme separation, only the component parallel to the baseline (the vector between reference and user) is not completely canceled by scalar DGPS. This is emphasized in Table 4.

Table 5 summarizes the residual errors attributable to satellite ephemeris after DGPS corrections are applied.

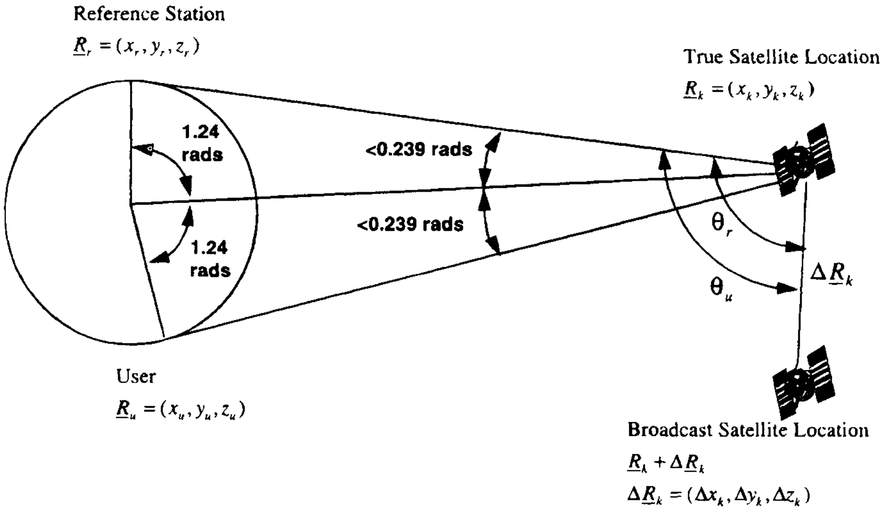

Reference Station
$\underline{R}_r = (x_r, y_r, z_r)$
True Satellite Location
$\underline{R}_k = (x_k, y_k, z_k)$
1.24 rads
<0.239 rads
<0.239 rads
1.24 rads
θ_r
θ_u
$\Delta\underline{R}_k$
User
$\underline{R}_u = (x_u, y_u, z_u)$
Broadcast Satellite Location
$\underline{R}_k + \Delta\underline{R}_k$
$\Delta\underline{R}_k = (\Delta x_k, \Delta y_k, \Delta z_k)$

Fig. 7 Worst-case separation of user and reference station.

The small type 2 error only occurs when there are large ephemeris errors. The expected velocity and accelerations shown in Table 5 are limited by the ephemeris message, which acts as a low-pass filter on the error, effectively limiting the magnitude of these effects.

In summary, SA manipulation of the ephemeris data in the navigation message could cause larger spatial decorrelation of the DGPS correction, but such manipulation is unlikely to cause meaningful temporal decorrelation. If SA is not applied to the ephemeris message, this is a negligible source of error for DGPS, provided the user is within 500 km of the reference station.

D. Ionospheric Errors for Differential GPS

Free electrons in the ionosphere produce a group delay in the GPS signal, which is a significant error source. The ionosphere is usually modeled as a

Table 4 Impact of ephemeris errors at maximum separation for scalar corrections

Satellite Ephemeris Component	DGPS error at maximal baseline separation (12,040 km) as a percentage of raw error
Radial (away from earth center)	0%
Parallel to baseline ("in plane")	47%
Perpendicular to above ("out of plane")	0%

Table 5 Residual differential GPS errors for satellite ephemeris

	Without DGPS Correction with or without SA Clock Dither		Zero Baseline Zero Latency DGPS (Type 3)		Decorrelation with Latency (Type 2)		Geographic Decorrelation (m/100 km)
	Bias (m)	Random (m)	Bias (m)	Random (m)	Vel. (m/s)	Accel. (m/s^2)	(Type 1)
Ephemeris Errors -SA not applied	10.0 extreme case	0.0	0.0	0.0	negl.	negl.	< 0.05
Ephemeris errors -SA applied to Ephemeris	100.0 extreme case	0.0	0.0	0.0	< 0.01	< 0.001	< 0.5

relatively thin blanket located at about 350 km above the Earth. Its effective vertical delay varies from a few meters in the early morning hours to 10–20 m at the maximum, which occurs about 2 h past local solar noon. This vertical delay must be multiplied by an "obliquity factor," which accounts for the angle with which the signal penetrates the blanket.

Under extreme conditions, the ionosphere can delay the satellite signal by many tens of meters because of the following: 1) solar storms during periods of solar maximum; 2) low elevation angles (high obliquity factor); or 3) peak delay conditions in the early afternoon.

More typically, vertical delays throughout a 24-h period are in the 4–10 m range. Without differential corrections, about 50–75% of this error can be removed by using a standard model and coefficients available in the navigation message (see Refs. 11 and 12, and Chapter 12 in the companion volume). A dual-frequency, P-code receiver can directly measure the delay and make a correction that should be accurate to about 1 m. As long as *both* or *neither* the user and reference station make a dual-frequency correction, the impact on DGPS should be errors of less than 1 m for separations of less than 100 km. It should be noted that long-range users have been successful in using differential ionosphere and troposphere models. These have reduced the geographic correlation. An example is discussed at the end of this chapter along with other test results.

Differential corrections for ionospheric delays will be in error because of the following.

1) The GPS signals received by the reference and user pass through ("pierce") the ionosphere blanket at different locations.

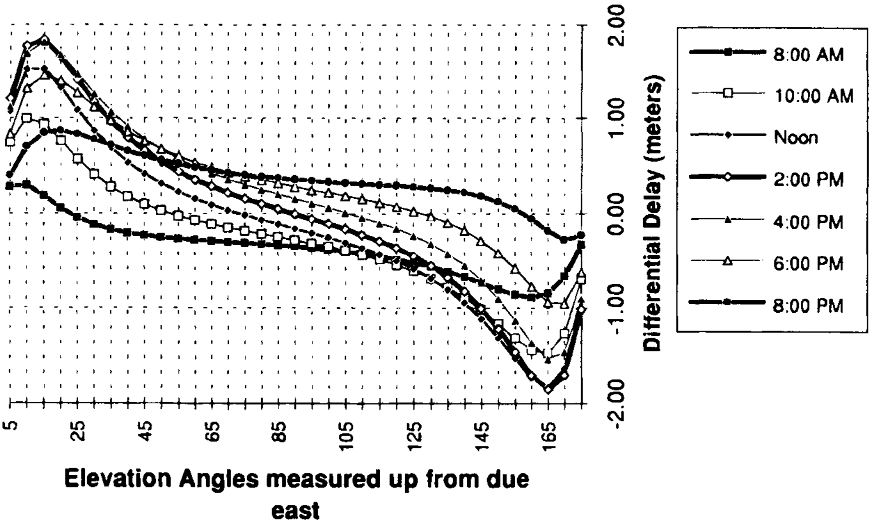


Fig. 8 The expected difference in ranging, m, attributable to the ionosphere for a 100-km separation due east. Each curve is for a different time of day at the reference station.

2) The incidence angle of the signal through the blanket is different (this is quantified by the obliquity factor).*

3) Latency provides outdated corrections (this is usually a smaller effect).

The impact of these effects is strongly a function of the time of day and has a small, relatively constant magnitude in the early morning hours.

1. Simulation of Ionospheric Decorrelations

With differential corrections, the size of the residual pseudorange error for the ionosphere depends most strongly upon the separation of the user and the reference station and the elevation angles of the satellites. Figure 8 predicts the size of this residual as a function of elevation angle and time of day. Larger separation distances will scale approximately linearly.

In this figure, the standard ionospheric model[12] is used to predict the signal delay at the reference station and at the user. As shown, as the elevation angle of the satellite decreases, the nominal ionospheric delay increases. If the user is assumed to be due east of the reference station, then the difference between the reference delay and the user delay also increases. Perhaps much of this residual delay can be modeled and removed (see test results in Sec. VII.B of this chapter); however, a residual error of $0.5 \times 10^{-6}|\Delta\bar{r}_{r,u}|$ to $5 \times 10^{-6}|\Delta\bar{r}_{r,u}|$ is expected (one sigma), where $|\Delta\bar{r}_{r,u}|$ is the reference station-to-user separation.

*The obliquity factor is the ratio of delays at any elevation angle to the vertical delay. It varies from 1.0 at 90 deg to about 3.0 at 5 deg. It is weakly a function of mean iononspheric height.

2. *Measured Ionospheric Decorrelations*

Ionospheric spatial decorrelation has been measured by Ref. 13, and these early measurements are summarized here. At ranges of 500 km, the residual errors were less than 1.8 m 95% of the time and less than 4.0 m 99% of the time. This effort to characterize differential residuals caused by the ionosphere is ongoing, and many years of data collection will be required for a complete characterization. However, these preliminary results suggest that the residual pseudorange error 1σ will be approximately $2 \times 10^{-6}|\Delta\bar{r}_{r,u}|$. As such, this spatial decorrelation is approximately equal to the decorrelation that would be introduced if the satellite ephemeris error were around 50 m.

3. *Summary of Ionospheric Errors for Differential GPS*

Under 50-km separation, the ionosphere is not a significant problem for dynamic DGPS systems. For the static surveyor, care should be taken beyond about 10 km, although the error (at two parts per million of range) is considerably better than a first-order survey. This is summarized in Table 6.

Table 6 Residual ionospheric errors for differential GPS[a]

	Without DGPS Correction with or without SA Clock Dither		Zero Baseline Zero Latency DGPS (Type 3)		Decorrelation with Latency (Type 2)		Geographic Decorrelation (m/100 km) (Type 1)
	Bias (m)	Random (m)	Bias (m)	Random (m)	Vel. (m/s)	Accel. (m/s^2)	
Ionospheric Errors (raw Ionosphere)	2 to 10 (times obliquity)	< 0.1 (times obliquity)	0.0	< 0.14	< 0.02	neglig.	< 0.2

[a]The values are typical 1σ estimates. The decorrelation with separation (between user and reference), although small, is the largest of the geographic decorrelation tersm.

E. Troposphere Errors for Differential GPS

The index of refraction of the lower atmosphere under standard conditions is not quite unity (it is typically 1.0003), and it depends upon temperature, pressure, and humidity. At low satellite elevation angles (below about 10 deg), tropospheric refraction can result in significant delays. Fortunately, most of this delay can be removed using a simple model that depends only upon satellite elevation angle and not on the local pressure, temperature, or humidity.

Without differential corrections, this model typically removes 90% of the delay, but the unmodeled error can reach 2–3 m at about 5 deg elevation. With differential corrections, the residual error is almost always very small. However, if the signal ray paths to the user and reference station traverse volumes with significantly different meteorological parameters, the error could be troublesome for demanding applications. For example, if the reference station and user are at significantly different altitudes (several thousand feet), then variations in the index of refraction could be significant. In these cases, the DGPS user should apply a differential tropospheric model that accounts for the altitude difference. These sensitivities are summarized in Table 7.

F. Local Area Differential GPS Error Summary

For convenience, the effects of various error sources on DGPS are summarized in Table 8. Two comments should be made. First, a poorly designed DGPS system will consistently be worse than these estimates of performance. Design deficiencies can occur in many elements, but the most common problems tend to be associated with the DGPS communications link. These are treated in more detail in Sec. VI of this chapter. Second, the table summarizes expected (one-sigma) values of pseudorange error. Because many of the underlying error sources are random, there will be times when they are better and times when they are worse.

Figure 9 shows the *ranging* error growth with latency *and* distance (it conservatively assumes that SA is applied to both clock and ephemeris). The *position* error suffered by a GPS (or DGPS) user is proportional to these pseudorange measurement errors, but it also depends upon the geometry of the user and the satellites. As discussed in Chapters 2, 5, and 11 of the companion volume, the measures that describe the degradation caused by geometry are known as dilution of precision (DOP) values. In fact, the position error is approximately equal to the relevant DOP value times the pseudorange error.

Many users are not comfortable with one-sigma values of pseudorange error and prefer the two-sigma values of positioning error that approximate the 95th

Table 7 Residual differential GPS user pseudorange errors caused by tropospheric delays

	Without DGPS Correction with or without SA Clock Dither		Zero Baseline Zero Latency DGPS (Type 3)		Decorrelation with Latency (Type 2)		Geographic Decorrelation (m/100 km)
	Bias (m)	Random (m)	Bias (m)	Random (m)	Vel. (m/s)	Accel. (m/s^2)	(Type 1)
Tropospheric errors	2 multiplied by obliquity	<0.1 multiplied by obliquity	0.0	<0.14	neglig.	neglig.	<0.1

Table 8 Summary of residual differential GPS pseudorange errors

	Without DGPS Correction		Zero Baseline Zero Latency DGPS (Type 3)		Decorrelation with Latency (Type 2)		Geographic Decorrelation (meters/100 km) (Type 1)
	Bias (meters)	Random (meters)	Bias (meters)	Random (meters)	Velocity (m/s)	Acceler. (m/s^2)	
Receiver Noise	0.5	0.2	0.5	0.3	0.0	0.0	0.0
Multipath	0.3 to 3.0	0.2 to1.0	0.4 to 3.0	0.2 to 1.0	0.0	0.0	0.0
Satellite Clock Errors	21.0	0.1	0.0	0.14	0.21	0.004	0.0
Satellite Ephemeris Errors S/A not applied	10.0 extreme case	0.0	0.0	0.0	negl.	negl.	<0.05
Satellite Ephemeris errors -S/A applied to Ephemeris	100.0 extreme case	0.0	0.0	0.0	<0.01	<0.001	<0.5
Ionospheric Errors (raw Ionosphere)	2 to 10 meters (times obliquity)	<0.1 (times obliquity)	0.0	<0.14	<0.02	neglig.	<0.2
Tropospheric errors	2 meters (times obliquity)	<0.1 (times obliquity)	0.0	<0.14	neglig.	neglig.	<0.2

percentile; that is, the position error that is expected to be exceeded no more than 5% of the time. This is called the 2drms value. An example for horizontal positioning is as follows:

$$\text{Horizontal 2drms} = 2\sqrt{\sigma_x^2 + \sigma_y^2} = 2\ \text{HDOP}\cdot\sigma_p$$

If the application requires three-dimensional positioning, then we use the following:

$$\text{Spherical 2drms} = 2\sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2} = 2\ \text{PDOP}\cdot\sigma_\rho$$

Because a typical value for horizontal *DOP* is 2.0, the 2drms horizontal position error for a user without differential corrections is about 85 m.* In contrast, the 2drms horizontal accuracy for a differential user within 50 km of the reference station varies from 1 to 5 m, depending upon the age of the correction and the quality of the system. A conservative LADGPS error budget is shown in Table 9.

*While this is driven by the SA-induced satellite clock errors that have been experimentally observed, there is no guarantee that it could not be larger (or smaller). The agreement with the U.S. Department of Defense (DOD) is that it will not exceed a 2drms horizontal error value of 100 m.

Next Page

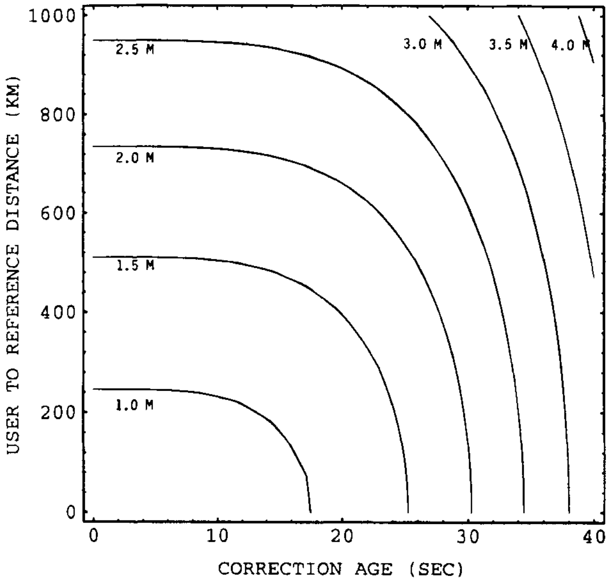


Fig. 9 **Growth in pseudorange errors from age (latency) and distance.**

Table 9 **Differential GPS error budget for users within 50 km of the reference station**

Error Source	1 sigma Error (meters)		
	Bias	Random	Total
Ephemeris Data	0.0	0.0	0.0
Satellite Clock	0.0	0.7	0.7
Ionosphere	0.0	0.5	0.5
Troposphere	0.0	0.5	0.5
Multipath	1.0	1.0	1.4
Receiver Measurement	0.0	0.2	0.2
Reference Station Errors	0.3	0.2	0.4
UERE (RMS)	1.0	1.4	1.8
Filtered UERE (RMS)	1.0	0.4	1.1
1 sigma errors--vertical	VDOP = 2.5		2.8
--horizontal	HDOP = 2.0		2.2

Chapter 2

Pseudolites

Bryant D. Elrod*
Stanford Telecom, Inc., Reston, Virginia 22090
and
A. J. Van Dierendonck†
AJ Systems, Los Altos, California 94024

I. Introduction

THE Global Positioning System (GPS) has reached operational status and is now being used internationally as a means for accurate, world-wide, all-weather positioning and navigation based on pseudonoise signals transmitted by a constellation of 24 satellites. Pseudolites (PLs) are ground-based transmitters that can be configured to emit GPS-like signals for enhancing the GPS by providing increased accuracy, integrity, and availability.[11,21]

Accuracy improvement can occur because of better local geometry, as measured by a lower vertical dilution of precision (VDOP), which is important in aircraft precision approach and landing applications. Accuracy and integrity enhancement can also be achieved by employing a PL's integral data link to support differential (DGPS) modes of operation and timely transmittal of integrity warning information. Availability is increased because a PL provides an additional ranging source to augment the GPS constellation.

Although the use of PLs offers many potentially significant benefits, a number of technical issues must also be addressed. One is the PL signal power level and the associated "near–far" problem that a user receiver may experience, depending upon the dynamic range of signal strength encountered as the distance to a PL changes. Other issues include deployment requirements, signal data rate, signal integrity monitoring, and user antenna location and sensitivity. This chapter addresses PLs from the perspectives of signal design considerations, integrated DGPS/PL considerations, and testing activities for assessing the reality and mitigation of various technical issues associated with using PLs.

*Vice President, GPS Navigation Systems.

†Systems Consultant.

II. Pseudolite Signal Design Considerations

Pseudolites operating within the GPS frequency bands (L_1: 1565–1585 MHz or L_2: 1217–1237 MHz) can be configured to serve a limited area with a power level low enough to preclude appreciable interference to standard GPS signals. One application is the PL bubble concept proposed for operating near the ends of airport runways to augment real-time kinematic positioning based on GPS carrier-phase measurements[3] (also see Chapter 15, this volume).

For PLs designed to cover a larger area, such as an entire airport or terminal area, potential interference to GPS signals is a key technical issue. Although operation outside the GPS bands is a possibility, this option ultimately adds rf front end complexity and cost to user equipment.

In this chapter, a PL signal structure is described that can operate within the L_1 band and mitigate or virtually eliminate the near–far issue. First, however, a brief review is given of previous implementations used on DOD test ranges and other approaches proposed for civil applications.

A. Previous Pseudolite Designs

Table 1 compares two DOD implementations and three other proposed approaches in terms of the diversity of techniques employed for interference mitigation: signal (code, frequency, time) and spatial separation. Pseudolites, called ground transmitters (GTs) during the Phase I GPS program, were used to augment GPS for testing user equipment at Yuma, AZ before there were enough satellites for navigation.[4] The PL signal structure was the same as used on the GPS satellites, except for the data content. In those days, interference to the satellite signals was avoided by using different gold codes (Code Division Multiple Access, or CDMA) and keeping the user equipment under test at an adequate distance from the four GTs to prevent dynamic range sensitivity problems. The same was true in more recent Space Defense Initiative (SDI) testing with GTs built by Stanford Telecom (STel) to augment GPS for DOD's Range Applications Program. However, a Time Division Multiple Access (TDMA) signal (on–off scheme) was also implemented to prevent interfering with a co-located GPS receiver that provided real-time synchronization to the GPS satellites. At transmit times, this receiver simply blanked out the signal.

The use of PLs for civil aviation applications was first proposed in 1984.[5] The "near–far" problem was recognized and addressed at that time, along with suggested solutions, but not carried any further. Three signal diversity techniques suggested as a solution were as follows:

1) Pulsed signals with random or fixed cycle rates, a TDMA variation

2) Signals transmitted at a frequency offset from L_1 (1575.42 MHz), but within the same frequency band as GPS, a variation of frequency division multiple access (FDMA)

3) Alternative codes that have a longer sequence than the existing GPS codes, a variation of CDMA

The first two alternatives were favored, with a preference toward the first. Although technically viable, the FDMA approach, including the possible use of L_2 (1227.6 MHz) or frequencies outside the GPS band, was considered more costly given the state of GPS receiver technology at the time. The CDMA

Table 1 Pseudolite interference mitigation techniques—previous designs

Pseudolite signal implementation (I) or recommendation (R)	Interference mitigation techniques			
	Code division multiple access	Frequency division multiple access	Time division multiple access	Pseudolite spatial separation
(I) Yuma ground transmitters for the Phase I GPS tests (1977)	Different gold codes	None	None	Unspecified
(I) Stanford Telecom ground transmitters for Space Defense Initiative (1989)	Different gold codes	None	Pulsed (simple on–off)	Unspecified
(R) Reference 5	Longer sequence codes	L_2 or L_1 ± 15 MHz	Pulsed at random or fixed rate	Unspecified
(R) Reference 6	Different gold codes	None	Pulsed with: 10% duty cycle, random pattern	≥ 30 km
(R) Reference 7	Different gold codes	$L_1 \pm \Delta f$ ($\Delta f \geq 30$ kHz)	Pulsed with: 1:11 duty cycle, fixed pattern, fixed offset between PLs	No constraint

approach with different pseudorandom noise (PRN) codes could be part of the diversity solution, but longer sequence codes would not add significant margin against cross-correlation interference.

As part of the Radio Technical Commission for Maritime (RTCM) user activities, a more definitive pseudolite signal structure was proposed in 1986.[6] All three of the above multiple access techniques were considered, but pulsed TDMA was the only approach recommended, because it made the least impact on the design of GPS receivers based on the state of technology at the time. Subsequently, flaws in that TDMA scheme were observed with respect to a class of "nonparticipating" receivers, some of which are still in use today. This led to a modified TDMA scheme, which was proposed in 1990.[7] Despite these proposals, fear of the near–far problem remained, and rightly so, because a limited interference margin can still exist with only code (CDMA) and pulsing (TDMA) employed.

B. New Pseudolite Signal Design

Fortunately, GPS receiver technology has advanced to a point that the FDMA approach is now viable, which improves the solution to the near–far interference issue significantly. As a result, a more effective signal structure has been proposed that combines: good C/A codes a frequency offset that takes advantage of the code cross-correlation properties, and a good pulsing scheme.[8]

1. C/A Codes

A recent study identified 19 C/A codes that were considered best (in terms of cross-correlation level) of the 1023 possible codes in the GPS C/A-code family.[9] These codes were selected for use with the FAA's wide area augmention system (WAAS), which will augment GPS with a ranging signal transmitted at L_1 from geostationary orbit. Although the 19 codes selected were the best, there remain 712 balanced codes, almost as good, that could be used for PLs. (See also Chapter 3 in the companion volume for additional discussion of Gold codes and cross-correlation properties.)

2. Frequency Offset

The proposed offset is 1.023 MHz on either side of L_1 at 1575.42 MHz, which places the PL carrier in the first null of the GPS satellite C/A-code spectrum. This offset is much more effective and one that at least some current GPS receivers can accommodate, because it resembles a large Doppler frequency offset. Furthermore, it virtually eliminates any code cross-correlation with the GPS satellite signals. This spectrum is shown in Fig. 1, which is a typical spectrum of a GPS C/A code [PRN 2]. It was reported years ago that the cross-correlation between C/A codes at different frequencies was simply the magnitude of another C/A-code spectral line at the offset frequency.[10] This property has since been verified. Because all the codes have similar spectral characteristics, the spectrum of any C/A code tells the story. As indicated in Fig. 1 and discussed further in Appendix A, the spectral lines near the null are well below −80 dB and are 60 dB lower than those at a zero frequency offset. This property has been used to

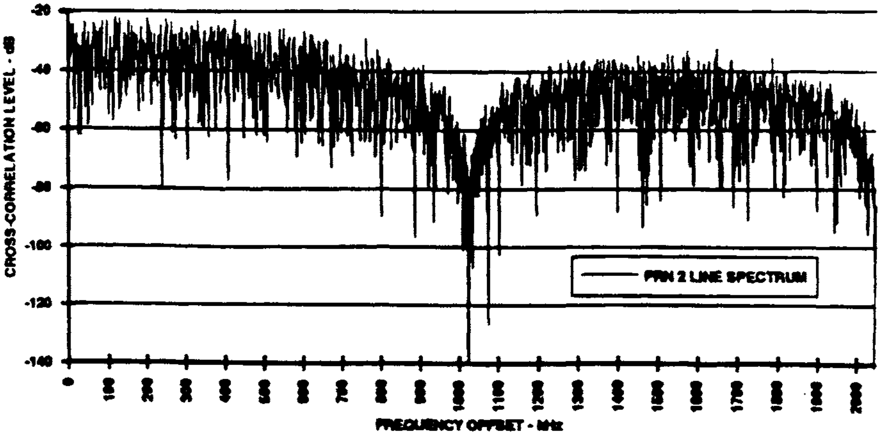


Fig. 1 Spectrum of the GPS C/A-code PRN 2.

advantage in the Global Orbiting Navigation Satellite System (GLONASS) signal structure.[11] All the GLONASS signals have the same maximum length codes with the first nulls at 511 kHz offsets from the carrier. The GLONASS FDMA signals are spaced at 562.5 kHz intervals, 51.5 kHz from the nulls of the adjacent signal.

Recently, it has been shown that cross-correlation levels can be somewhat higher than indicated above when the code chip boundaries are not aligned.[12] However, when the carrier frequency is offset by 1.023 MHz, and the carrier/code frequency ratio of 1540 is maintained, the code of the PL is shifting with respect to the satellite codes at a rate of over 664 chips per second. Thus, any cross-correlation is noise-like and averages to a lower rms level. As pointed out, this is still interference.[12] An in-band frequency offset of 1.023 MHz lowers the rms interference by about 8 dB, but does not eliminate it. However, it does eliminate cross-correlation problems.

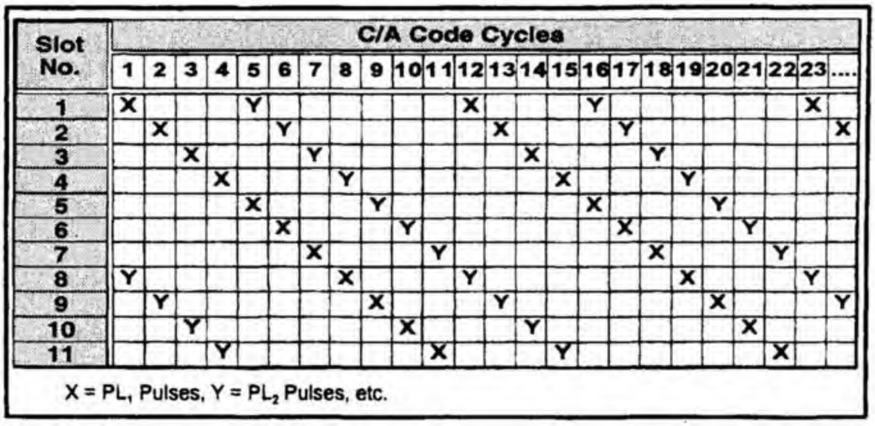
3. Pulsing Scheme

The FDMA approach described above is a variation of an earlier proposal, which suggested a continuous signal with a longer code on the fringe of the allocated L_1 band at 1560 or 1590 MHz. However, a *continuous* in-band PL signal could still cause interference problems, depending upon its power level compared to GPS satellite signals. For example, a PL signal whose power is set to be received at 37 km (20 n.mi.) will be 60 dB stronger at 37 m (0.02 n.mi.), which will be approximately 30–40 dB above the noise. Thus, although offsetting the frequency is a good idea, TDMA pulsing may still be required to avoid this situation. Furthermore, if TDMA is used, there is no reason to operate on the fringe of the L_1 allocated band, because that large a frequency offset and a longer code are both undesirable due to the additional burden on the GPS receiver, even with modern technology.

With a good pulsing scheme, the impact on the reception of GPS signals can be made essentially transparent. The receiver treats it as a continuous signal, provided that it is designed to suppress pulse interference, as most modern receivers are, even if by accident. As noted in the RTCM work,[6] any hard-limiting receiver or "soft-limiting" receiver will clip the pulses and limit their effect, but still pass more than enough pulse power to track the PL signal itself. A "soft-limiting" receiver clips the incoming signal-plus-noise at two to three times the rms noise level, resulting in more, but still negligible degradation in signal-to-noise performance.

All modern digital receivers are either hard-limiting or possess the soft-limiting property through precorrelation quantization. Although this is not true for the older, analog military receivers, their wideband automatic gain control (AGC) suppresses the pulses for the same effect. Any cross-correlation problems in those receivers caused by PL transmissions are eliminated with the proposed frequency offset.

The pulsing scheme presented here differs from previous proposals.[6,7] It accounts for the fact that the PL message symbol rate could be 50 N sps (where N = 1, 2, 4, 5, 10, or 20). The recommended pulse pattern is illustrated by the Xs shown in Fig. 2. These repeat every 11 ms, and thus, would never be synchro-



Slot No.	C/A Code Cycles																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	....
1	X				Y							X				Y							X	
2		X				Y							X				Y							X
3			X				Y							X				Y						
4				X				Y							X				Y					
5					X				Y							X				Y				
6						X				Y							X				Y			
7							X				Y							X				Y		
8	Y							X				Y							X				Y	
9		Y							X				Y							X				Y
10			Y							X				Y							X			
11				Y							X				Y							X		

X = PL$_1$ Pulses, Y = PL$_2$ Pulses, etc.

Fig. 2 Illustration of pseudolite pulse pattern.

nous with a received GPS bit pattern. Each code cycle (1 ms) is divided into 11 slots, each with a width of 1/11,000 s (90.90909 μs). The pulses would never be synchronous with a received GPS bit pattern, as would occur with the RTCM design.[6]

Only one of these slots contains a pulse, so the duty cycle is 1/11. There will be 20/*N* pulses per symbol, but only one is required because the receiver would integrate the energy over the entire symbol period. Because every slot is filled once every 11 ms, the entire C/A code would be received every 11 ms. The clocking rate for the slots is 1/93 of the C/A-code chipping rate of 1.023 MHz, or 11 kHz. It is noted in Appendix B that with a duty cycle of 1/11, the loss of GPS satellite signal-to-noise ratio, in either an analog receiver with pulse suppression, or a digital receiver with natural *soft-limiting,* is less than 1.5 dB when close to the PL.

Potential mutual interference when multiple PLs are installed in an area was reported in the RTCM work.[6] With the RTCM pulse pattern, a minimum distance between PLs would be required, because the transmission of pulses simultaneously from each PL could result in the simultaneous reception of multiple pulses. But, that need not be, because the pulse timing of multiple PLs can be offset. Unfortunately, the RTCM pattern had irregular times between pulses, so that there would still be some simultaneous receptions.

However, with the pulse pattern shown in Fig. 2, a suitable pulse-timing offset would prevent simultaneous reception from ever occurring, except at large distances where the PL signal would be of little consequence. For example, consider two PLs where the pulse timing was offset by 4 ms as indicated by the Xs and Ys in Fig. 2. Because the minimum transmit time separation between X and Y pulses is 4/11 ms, only receivers with a differential range more than 110 km from the PLs would encounter simultaneous reception. At that distance, the received powers of either PL would be negligible.

4. *Pseudolite Carrier Tracking*

A potential misconception about pulsing a PL signal is that it will prevent the tracking of PL carrier phase. This is not true if the pulses occur at a high enough rate, which is the case in the proposed scheme. One or more pulses will always be integrated along with noise over each symbol. The result is transparent to the tracking loops, because the phase change due to Doppler uncertainty over a symbol is negligible. To the tracking loops, it looks like a continuous signal.

5. *Pseudolite Transmit Power*

Given an average received power P_r through a receiving antenna (with gain G_a) at a distance d (in n.mi.) from a PL, the average transmitted power (P_T) at $L_1 \pm \Delta f$ is as follows:

$$P_T \approx P_r + 20 \log_{10}\left(\frac{4\pi d}{\lambda_1}\right) - G_a$$

$$= P_r + 20 \log_{10} d + 101.75 - G_a$$

where $\lambda_1 = 0.00010275$ n.mi. is the signal wavelength corresponding to L_1. As an example, consider an average received power of -130 dBm at 20 n.mi. (37 km) through an antenna gain of -10 dB (assumed for a small negative elevation angle). The average transmitted power is 7.77 dBm, or about 6 mW. The peak power for a duty cycle of 1/11 is then 18.18 dBm, or about 66 mW.

III. Integrated Differential GPS/Pseudolite Considerations

The introduction of PLs has two key objectives: signal augmentation and data link enhancement. The first is to increase the number of available signals and, thereby, improve or maintain the geometrical quality for user position determination. The second is to provide an integrated capacity for supplying key data to users for GPS (and PL) integrity warning and differential corrections to improve positioning accuracy via code-based local differential GPS (LDGPS) and potentially carrier phase-based kinematic differential GPS (KDGPS) techniques.

Although the definition of a compatible signal format is one of the most critical requirements for meeting these objectives, there are other implementation-related aspects that need consideration as well. This section addresses five of these: PL siting, PL time synchronization, antenna location on a user aircraft, the PL data message, and the filter algorithm for integrated GPS/PL measurement processing.

A. Pseudolite Siting

It is well known that the GPS geometry (quantified in terms of HDOP and VDOP factors) will vary over time and user location, even with a full 24-satellite constellation[13] (also see Chapter 5, the companion volume). It is also known that at times significant degradation (VDOP >> 6) can occur if fewer satellites are active. The utility of PLs for geometric enhancement lies in the fact that lower DOP values with less temporal variation can be achieved.

Figures 3 and 4 illustrate representative HDOP and VDOP profiles determined for a full constellation[13] and with one satellite inactive for cases where the user employs the best four GPS satellites or "all-in-view" with augmentation by one or two PLs.[2] In this illustration, the user was assumed to be at an altitude of 200 ft (the decision height for a Cat. I approach) with the PLs located 1 n.mi. ahead and/or 2 n.mi. behind. It is evident that in this situation, a two-PL augmentation would enable HDOP $\leq$ 1.0 and VDOP $\leq$ 1.5 *continuously.* At those levels, code-based LDGPS with the capability to correct uncorrelated pseudorange errors to $\leq$ 1m (2σ) would meet the current Cat. III (horizontal) and Cat. II (vertical) sensor accuracy requirements of 4.1m and 1.4m (2σ), respectively for aircraft precision approaches.[14] With the addition of ranging-capable satellites in geostationary orbit (e.g., via the FAA's WAAS implementation) to offset one or two inactive GPS satellites an even lower VDOP is possible with favorable implications for meeting the most stringent Cat. III (vertical) accuracy requirement of 0.4m (2σ).[14]

An alternate approach to meeting the Cat. III vertical accuracy requirement for precision approaches is based on KDGPS techniques with PL augmentation. Recent flight tests have demonstrated that a varying user/PL geometry will permit rapid resolution of carrier-range ambiguities to provide KDGPS-based measurements with centimeter-level precision[3] (also see Chapter 15, this volume). The configuration adopted for this technique utilizes a *bubble* concept in which two PLs located ahead of the runway threshold on each side of the glideslope transmit a standard, unpulsed L_1-C/A signal at a low power level sufficient to be received within a hemispherical bubble centered at each PL. A descending aircraft would enter the common bubble zone, acquire the PL signals, resolve GPS carrier-phase ambiguities, and emerge to continue the approach with KDGPS-based corrections supplied via a separate (non-PL) datalink. As proposed, four active PLs per runway would be needed to support landings in either direction.

On the other hand, wider coverage *large bubble* PLs with interference mitigation inherent in the signal design could provide service to multiple runways over an entire airport region. Location of the PL antenna off runway and at suitable elevation would also be desirable from the standpoint of user antenna requirements (discussed in Sec. II.C). Pseudolite-siting requirements (number and location) to accommodate multirunway situations, user antenna considerations, and local constraints are key issues needing further analysis and testing.

B. Pseudolite Time Synchronization

Synchronizing a PL's clock to GPS time can be achieved in two ways—one in which a PL is collocated with an LDGPS reference receiver (RR), and one in which it is remote from the RR that is tracking its transmitted signal. In the latter case, the RR sends corrections to the remote "slave" PL to correct itself and also sends message data (e.g., code- and/or carrier-based DGPS corrections) to be transmitted. These two different PL configurations are illustrated in Fig. 5. In the co-located configuration, the RR shares the transmit/receive antenna with the PL, which also allows for self-calibration.

The type of configuration used would depend upon whether or not there is more than one PL at a given local region. If there is only one, the collocated

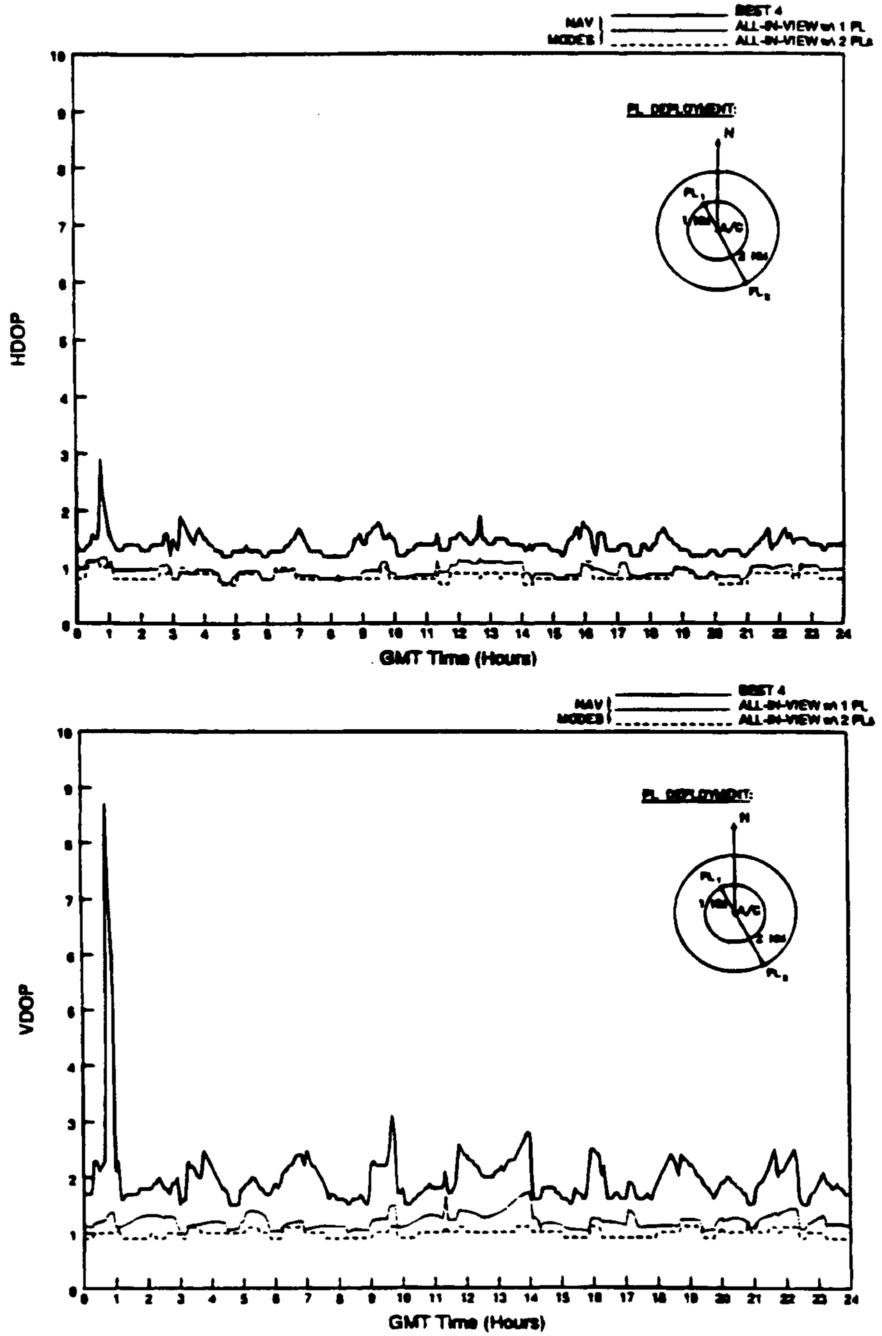


Fig. 3 Worst-case dilution-of-precision profiles at 35°N with and without PLs for a full (24) GPS constellation.

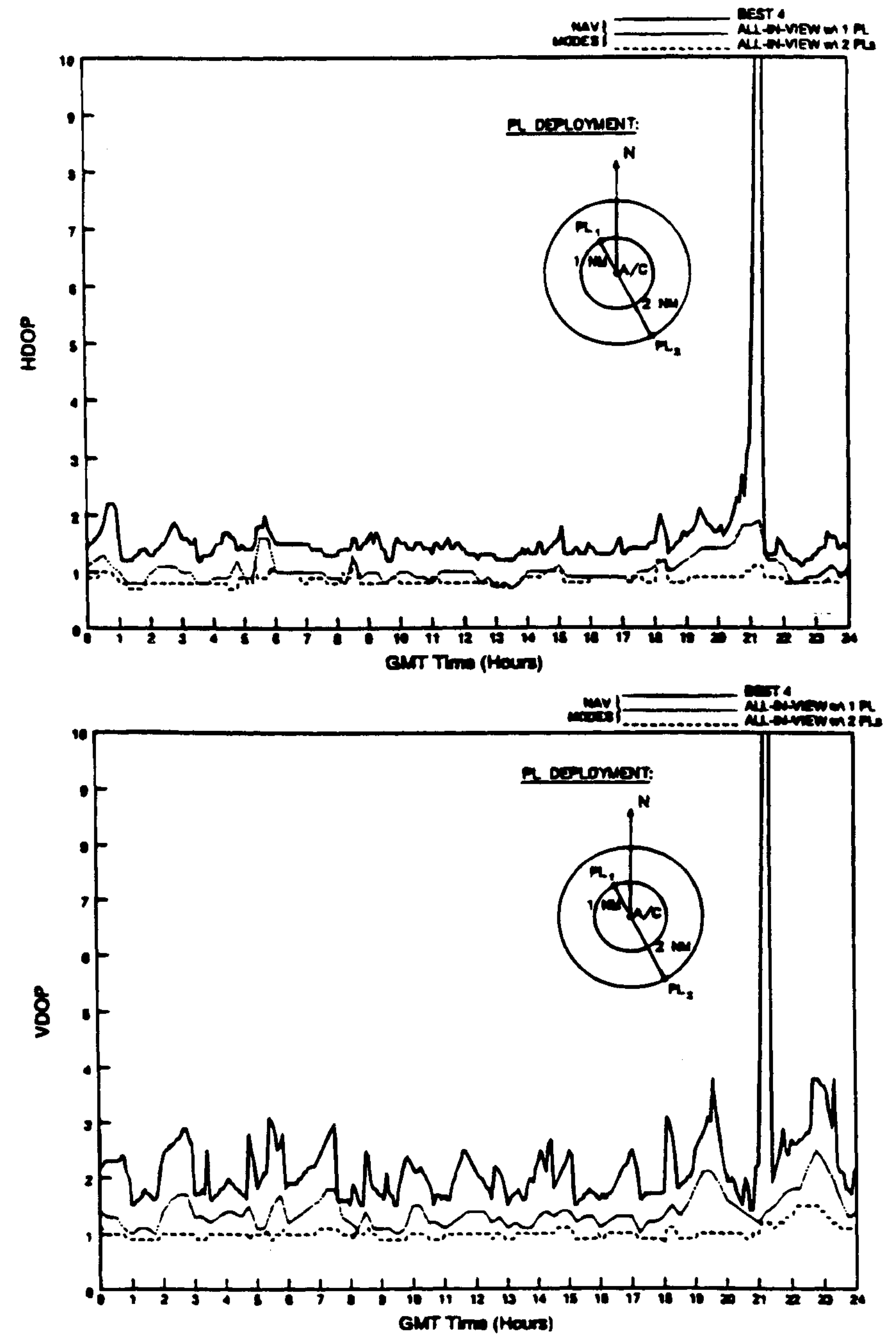


Fig. 4 Worst-case dilution-of-precision profiles at 35°N with and without PLs for a GPS constellation with one inactive satellite.

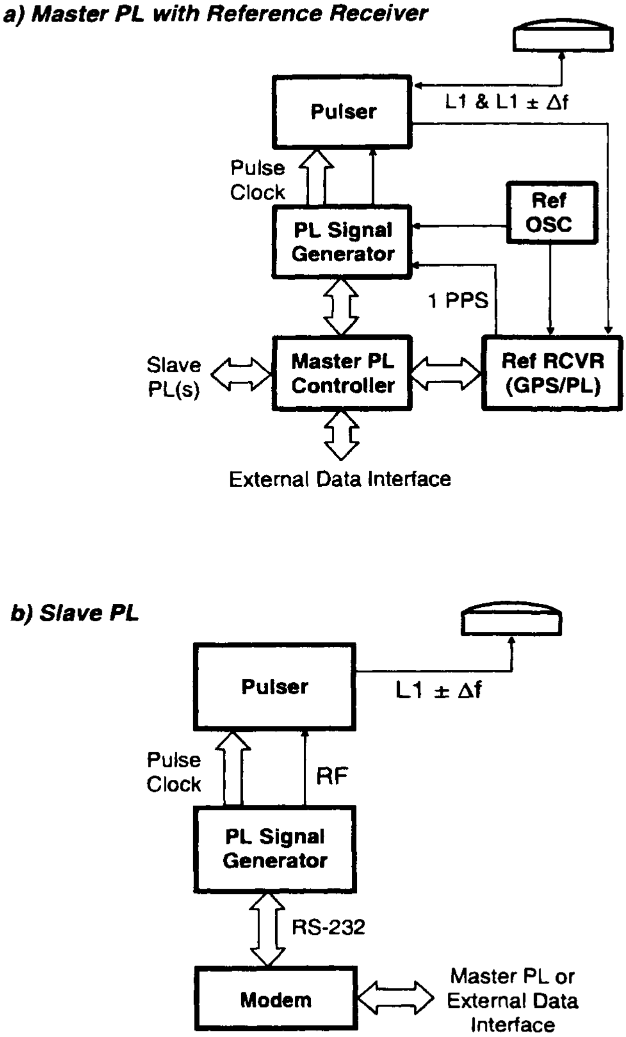


Fig. 5 Master and slave pseudolite configurations.

approach is more desirable, especially if line-of-sight (LOS) visibility to the RR might be a problem. If there is more than one PL, the remote approach with a common RR for synchronization is more desirable. However, if LOS visibility problems exist, having a RR collocated with each PL would enable time synchronization via GPS common-view, time-transfer techniques. However, only one RR can be used for deriving satellite LDGPS corrections. In this case, master and slave PLs would be designated with the master clock and LDGPS corrections derived centrally and distributed to each remote PL, which would slave its time to the master RR clock. This approach is similar to that used by various DOD test and training ranges coordinated by the Range Applications Joint Program Office (RAJPO), for which P-code PLs were developed by Stanford Telecom.

1. Master Pseudolite Configuration

Both the RR and the PL signal generator derive their timing coherently from the same stable frequency standard. The signal generator/pulser electronics module pulses the transmission of the PL signal in order to minimize interference to both participating and nonparticipating GPS receivers, as discussed in Sec. IB3. This multiplexing also allows the RR to receive the satellite signals via the same antenna. In fact, by providing a suitable calibration path, the RR can also track the output of the signal generator. In this way, the collocated PL is self-calibrating, and the transmitted PL signal will be synchronized to the same clock that is used to derive the differential GPS corrections. This is true even in the multiple-PL scenario, where the slave PLs receive differential corrections from the master PL. In this case, the time solutions of the slave PL receivers will be referenced to the master PL's clock, because the differential corrections are computed with respect to that clock.

2. Slave Pseudolite Configurations

The slave PLs need not have receivers if they can be tracked by the RR. The RR simply supplies corrections to the PL for correcting its clock via a number-controlled oscillator (NCO) and provides the satellite differential corrections for modulation on the PL signal. Because the RR can update the PL continuously, a good quality crystal oscillator will suffice as its frequency reference. Otherwise, its configuration is identical to the co-located configuration but without the RR and the self-calibration path.

C. User Aircraft Antenna Location

Reception of the PL signal by a user will be affected by the aircraft antenna location and the PL position relative to its approach path. Ideally, the PL signals would be received by a top-mounted antenna, the same one used for receiving the satellite signals. This could be accommodated by locating the PL antenna at an appropriate elevation and offset from the glideslope. Even if the line of sight to the PL is below the aircraft antenna horizon, the increased signal level in the vicinity of the PL will tend to cancel the loss in antenna gain.

If the aircraft were to pass directly over a PL, a larger angular gradient would be available to support the positioning process. However, a bottom-mounted antenna and a separate front end to interface the antenna(s) to the receiver may be needed. A configuration with a bottom-mounted antenna would likely have more sensitivity to ground-reflected multipath and interference. A top-mounted antenna is less sensitive, because reflective surfaces on the aircraft are typically small relative to the C/A-code chip width (293 m). In addition, the use of a bottom-mounted antenna would introduce additional sensitivity to lever arm uncertainty and knowledge of aircraft attitude information. On the other hand, it could be used directly, if a PL were to be employed only as a datalink. A bottom antenna *is* required for the small bubble concept.[3] Obviously, further analysis and testing are needed before the aircraft antenna location issue is resolved, one way or the other.

D. Pseudolite Signal Data Message

A PL offers the possibility for an order-of-magnitude increase in the data rate (up to 1000 bps vs 50 bps for GPS) that can be supported via a GPS-compatible signal with essentially a firmware change in the user receiver. Validation of data link performance at this rate is a key test objective. A closely related issue is the PL message capacity required to support GPS and PL integrity updates and DGPS corrections (code and/or carrier).

Some have proposed that for the latter, a 2400 bps data link would be needed to supply raw carrier-phase and other GPS information from a reference receiver at a 0.5–1.0 Hz update rate. Efforts by RTCM SC-104 (Special Committe 104) and others contend that KDGPS could be supported with much lower data transfer requirements ($\leq$ 1 kbps) based on the use of carrier phase *corrections,* not raw data.[15] A related issue for testing is whether PL aiding of carrier-phase ambiguity resolution and cycle slip maintenance procedures could reduce the data requirements still further.

1. General Format

For efficiency, the general format is patterned after the WAAS format[16] with three differences—it may or may not include forward error correction (FEC), a 7-bit distributed time word is added and the data rate can be higher than 250 bps, up to 1 kbps without FEC. This general PL format is shown in Fig. 6. The time word is added as a convenience to the user receiver, because PL time is not the same as GPS time. This is because the code-chipping rate is offset in frequency by approximately ± 664 chips per second, the feature that eliminates cross-correlation with the GPS signals. The actual PL frequency can be chosen (at an offset from the spectral null of 710.4166667 Hz) so that the PL week is exactly 393 s shorter or longer than the GPS week. In fact, PL time is, at any GPS time, as follows:

$$t_{\mathrm{PL}} = \left(1 \pm \frac{393}{604{,}800}\right) t_{\mathrm{GPS}}$$

depending upon which null is selected. The PL time is distributed over three 250-bit subframes making up a 21-bit word. The 21-bit time word represents the

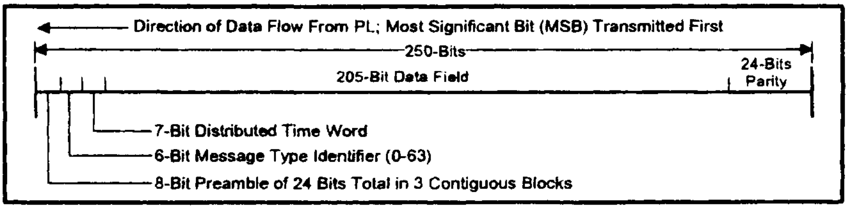

Direction of Data Flow From PL; Most Significant Bit (MSB) Transmitted First
250-Bits
205-Bit Data Field
24-Bits Parity
7-Bit Distributed Time Word
6-Bit Message Type Identifier (0-63)
8-Bit Preamble of 24 Bits Total in 3 Contiguous Blocks

Fig. 6 Pseudolite message format.

PL time of week at the start of the currently transmitted 24-bit preamble, starting with 0 at the beginning of the week. This PL time word also serves as the reference time for the data in the messages.

The 24-bit parity is the same as the cyclic redundancy check (CRC) parity specified for the WAAS.[16] The data field consists of 205 information-bearing bits.

2. *Message Types*

To avoid confusing the message types with those of the WAAS, the message type numbers start at 40. Table 2 lists a tentative set of PL message types. Every message would include a certain number of GPS/PL signal integrity flags to provide a short time to alarm capability.

3. *Message Content*

The contents of the messages are somewhat different from the message contents defined for the RTCA SCAT-I (Special Category I) DGPS.[17] This is to accommodate the 205-bit data fields and to provide data that are more consistent with Category II and III precision approach requirements. The SCAT-I messages are quite inefficient in requiring too much signal bandwidth. Preliminary message contents show more than adequate bandwidth in using the maximum 1 kbps capability of the proposed PL signal. Every message broadcast contains integrity flags for several PRN numbers, with a minimum of 11, including the transmitting PL. This allows for a positive integrity indication at least once per 0.5 s. PRN numbers that accompany the flags do not have to be broadcast in order, so an alarm can always be inserted in any 0.25 s message providing a maximum time to alarm of 0.5 s.

Message Type 42 consists of a slight modification to the content of the corresponding RTCM carrier-phase corrections message.[15] Each can accommodate four satellites/pseudolites. Therefore, both pseudorange and carrier-phase corrections can be broadcast.

E. GPS/Pseudolite Navigation Filter Considerations

When pseudorange (or carrier-range) measurements are processed by the user navigation filter, modeling to accommodate the nonlinearity in the measurement

Table 2 Pseudolite message types

Type	Contents
40	Don't use this PL for anything (PL testing)
41	Integrity flags/Pseudorange corrections
42	Integrity flags/Carrier-phase corrections (if required)
43	Integrity flags/PL location and PRN assignment
44	Integrity flags/PL almanacs
45	Integrity flags/Precision approach path definition
46	Integrity flags/Special message
63	Null message—Alternating 1s and 0s

model can be ignored in the case of satellites, but not necessarily for PLs. As the user range to the PL diminishes during an approach, the impact of the nonlinearity is to introduce an apparent bias into the measurements. If this error is comparable to the measurement error, then a standard extended Kalman filter (EKF) will yield inferior performance. Filter divergence may occur, as the combination of measurement and nonlinearity error exceeds the filter's own computation of rms error, and it rejects new measurement data.

One approach to preventing filter divergence of this sort is to choose sufficiently large a priori measurement variances that include worst-case nonlinearity effects. Unfortunately, this requires identification of what *is* worst case. More importantly, high constant measurement variances will cause sluggish performance at other times when there is a negligible nonlinear effect present.

Another approach, as outlined in Appendix C, is to increase the filter sophistication by using a Gaussian second-order (GSO) filter that is similar to a (linearized) EKF but includes a quadratic component of the measurement nonlinearity. The key benefit is that it offers improved performance when the nonlinear effect is present but reverts naturally to a standard EKF when it is not. A possible downside is that more software and processing time are required, although some approximations are possible depending upon the scenario. Pseudolite siting relative to the approach trajectory will also be a factor in this.

IV. Pseudolite Testing

The PL concept for augmenting GPS, as described above, is intended to provide users with the following potential benefits: 1) enhanced local area navigation performance via integrated GPS/PL positioning using single or multiple PL signals with airportwide or even terminal area coverage depending upon the power level and antenna configuration (s) employed; 2) more data link capacity to support DGPS operations directly (code and/or carrier-based) at a multiple N of 50 bps (where $N = 2, 4, 5, 10,$ or 20); and 3) mitigation of potential interference to standard GPS signals through the code, frequency, and time diversity techniques employed in the PL signal design.

Initial testing to verify these features has been conducted under a research and development project sponsored by the FAA,[18] and further comprehensive testing is planned. The following subsections describe the test results.

A. Pseudolite Interference Testing

Initial laboratory tests to evaluate PL interference impacts on GPS receivers was conducted by applying simulated GPS and PL signals at various power levels to several commercial receivers. After initial signal acquisition and steady-state operation by the GPS receiver was reached, a PL signal was injected at gradually increasing power levels. Signal quality reported in terms of GPS carrier power-to-noise power density ratio (C/N_0) was recorded as a function of average PL signal power. The plots in Fig. 7 show the test results obtained for two receivers under the following PL signal conditions: 1) no pulsing, no frequency offset; 2) no pulsing, frequency offset applied; 3) pulsing applied, no frequency offset; and 4) pulsing and frequency offset applied.

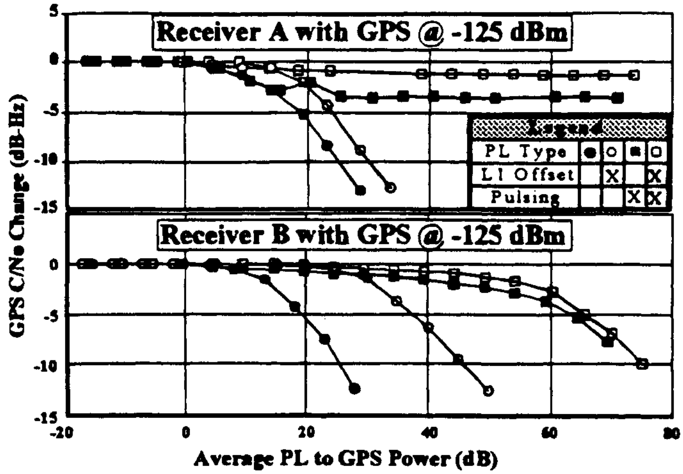


Fig. 7 GPS degradation vs pseudolite signal power with and without pulsing and frequency offset.

With no pulsing or frequency offset applied, the degradation is significant at just a 20 dB advantage in PL over GPS signal power. With either pulsing or frequency offset applied, the degradation is less, although pulsing is the more effective feature. With both applied, the degradation in reported C/N_0 from the *no PL signal* case is small ($\leq$ 2 dB) over a range of 60 dB or more in the ratio of PL to GPS signal power. The results for this case from testing four receivers (two L_1-C/A only and two L_1/L_2 cross-correlating types) are shown in expanded scale in Fig. 8. When compared to a user receiving a PL signal at a range of 20 n.mi. (37 km) at a level comparable to GPS, this means that a degradation of only 2 dB would be experienced by a user only 0.02 n.mi. (37 m) away but receiving the PL signal at 60 dB higher power!

Note that receiver A's degradation with a continuous signal, or with either a frequency offset or pulsing, but not both, is more than that of receiver B. This

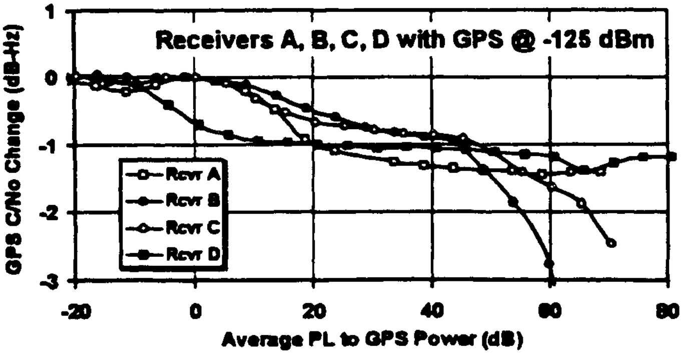


Fig. 8 GPS degradation vs pseudolite signal: 4 receivers with pulsing and frequency offset.

illustrates the difference between *soft-limiting* and *hard-limiting*. Receiver A uses multibit sampling. However, with both features turned on, receiver A's performance is quite good. Note also that receiver B's performance tails off as the PL signal gets quite strong (> 50 dB above GPS). Receiver B is a discontinued model, and receiver C is its replacement. Receiver C did not exhibit the same behavior, which was probably because of slow saturation recovery in the receiver B's front end. For receivers with fast enough saturation recovery, the results agree well with Eq. (B12) and Eq. (B13) of Appendix B.

The PL interference test results discussed above were obtained with *high-end* receivers capable of reporting signal quality data. Testing of other relatively low-cost commercial receivers was also done on a qualitative basis. These receivers were placed at varying distances from an antenna radiating a PL signal with pulsing and frequency offset features applied or not. With the PL signal off, each receiver was set up for normal GPS operations. With the PL signal on at peak power, each receiver was moved toward the PL antenna, and the separation distance was recorded when anything anomalous appeared on its display. The results for the four units tested in this manner showed no effect until within 1–5 m of the PL antenna at a peak radiated power of +15 dBm.

B. Pseudolite Data Link Testing

Tests were performed to verify the capability to transmit, receive, and demodulate the message data encoded on a PL signal with the pulsing, frequency offset, and higher data rate features described above. During laboratory tests at Stanford Telecom, a signal generator/pulser unit was interfaced directly to a GPS/PL-capable receiver. Pseudolite signals with a fixed data message were transmitted at different power levels to simulate operations at various PL/user ranges. Results of these initial tests indicated virtually error-free data reception, with the received signal quality (C/N_0) at a level of 35 dB-Hz or more.

To support planning for more comprehensive PL testing and evaluation, preliminary flight trials of a DGPS data link provided by one PL were arranged. A GPS/PL reference receiver and the signal generator/pulser unit used for the laboratory tests were installed at the FAA Technical Center (FAATC) and configured to transmit the PL signal from an antenna on the hanger roof. The test aircraft was an FAA Aerocommander (N50) with a top-mounted, low-profile GPS antenna, a GPS/PL-capable receiver, and data processing/storage equipment. For test purposes, the PL message data (LDGPS) corrections) were encoded in a 250-bit WAAS message format and transmitted without forward error correction at 250 bps.

Flight profiles flown with the test aircraft included straight-in approaches to Runway 6 at FAATC and various holding patterns ranging up to 29 km (15 n.mi.) from the PL. Received PL message data and signal quality data (C/N_0) were recorded onboard during flight segments, such as A, B, and C in Fig. 9.

Corresponding C/N_0 profiles and occurrences of individual data message errors are shown in Fig. 10. The PL data error occurrences appear to be predominantly associated with PL antenna gain reduction/obstruction phenomena during turns. Future flight trials will help address the impact of aircraft antenna type on PL signal-tracking performance. The test aircraft will be equipped to receive PL

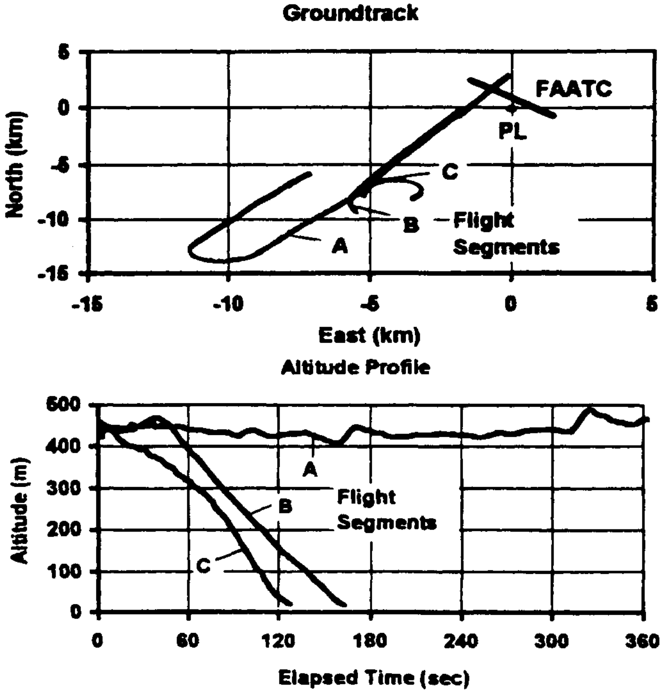


Fig. 9 Examples of flight test segments at the FAA Technical Center during pseudolite data link tests.

signals from top- and bottom-mounted, low-profile antennas, and a high-profile (blade) antenna. These trials will also assess PL data link reliability at other data rates (500 and 1000 bps).

C. Navigation Performance Testing

The initial laboratory tests and preliminary flight testing have focused on demonstrating the feasibility of PL tracking and interference mitigation to GPS signals with the pulsing and L_1 offset features applied. Additional aspects that remain to be assessed include the following:

1) PL code and carrier measurement quality
2) accuracy performance enhancement in LDGPS/PL and KDGPS/PL modes
3) sensitivities to PL siting (number/location) and user antenna (type/location)

Future flight trials are planned to assess navigation performance in LDGPS/PL and KDGPS/PL modes, and the sensitivity to PL siting and type/location of user antenna employed.

Independent KDGPS-based tracking of the test aircraft will be used as a truth source for assessing the consistency of PL ranging data (code and carrier) throughout a terminal area flight envelope.

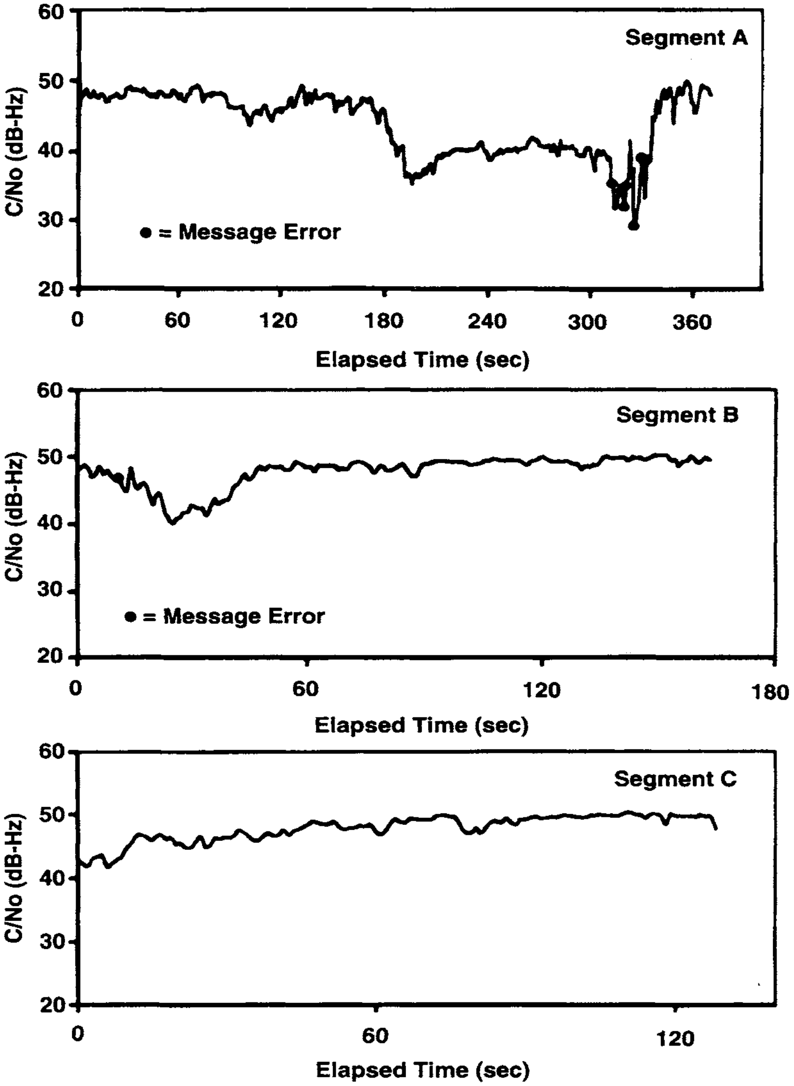


Fig. 10 Pseudolite signal profiles during data link tests.

Appendix A: Interference Caused by Cross Correlation Between C/A Codes

The effect that one C/A code has on another with respect to cross-correlation properties was described in Ref. 10 (see also Chapter 3, the companion volume). A more rigorous derivation is provided here based on the receiver signal-processing model shown in Fig. A1.

In this signal-processing model, the input signals at ①, in terms of in-phase and quadraphase components, are the following desired signal:

$$I_{i1}(t) = A_i C_i(t)\cos(2\pi\Delta f_i t + \phi_i)$$
$$Q_{i1}(t) = A_i C_i(t)\sin(2\pi\Delta f_i t + \phi_i) \tag{A1}$$

and the undesired signal

$$I_{j1}(t) = A_j C_j(t + \tau_j(t))\cos[2\pi(\delta f_j + \Delta f_j)t + \phi_j]$$
$$Q_{j1}(t) = A_j C_j(t + \tau_j(t))\sin[2\pi(\delta f_j + \Delta f_j)t + \phi_j] \tag{A2}$$

where

A_i, A_j = signal amplitudes

$C_i(t), C_j(t)$ = signal C/A codes

$\Delta f_i, \Delta f_j$ = signal Dopplers

ϕ_i, ϕ_j = signal phases

δf_j = frequency offset of undesired signal

$\tau_j(t)$ = time offset between signals

The Doppler removal process eliminates the desired signal's Doppler and phase so that the signal components at ② in Fig. A1 are as follows:

$$I_{i2}(t) = A_i C_i(t)$$
$$Q_{i2}(t) = 0 \tag{A3}$$

$$I_{j2}(t) = A_j C_j(t + \tau_j(t))\cos[2\pi(\delta f_j + \Delta f_j - \Delta f_i)t + \phi_j - \phi_i]$$
$$Q_{j2}(t) = A_j C_j(t + \tau_j(t))\sin[2\pi(\delta f_j + \Delta f_j - \Delta f_i)t + \phi_j - \phi_i] \tag{A4}$$

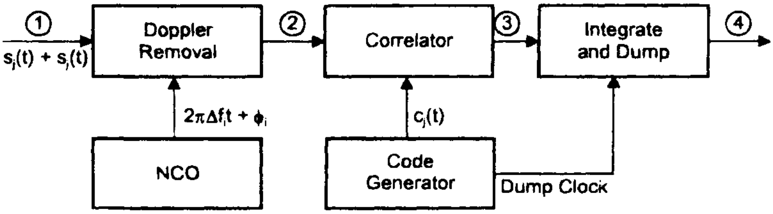


Fig. A1 Receiver signal-processing model.

where the Doppler difference is as follows:

$$\Delta f_{ij} = \delta f_j + \Delta f_j - \Delta f_i \tag{A5}$$

For simplicity, assume that the code transitions line up. Thus, for the moment, assume the following:

$$\tau_j(t) = NT_c \tag{A6}$$

where N is an integer, and T_c is a chip width (1/1,023,000 s). Also, assume full correlation for signal i. Then, at ③ we have the following:

$$\begin{aligned} I_{i3}(t) &= A_i \\ Q_{i3}(t) &= 0 \end{aligned} \tag{A7}$$

$$\begin{aligned} I_{j3}(t) &= A_j C_j(t + NT_c) C_i(t) \cos(2\pi \Delta f_{ij} t + \phi_j - \phi_i) \\ Q_{j3}(t) &= A_j C_j(t + NT_c) C_i(t) \sin(2\pi \Delta f_{ij} t + \phi_j - \phi_i) \end{aligned} \tag{A8}$$

Now, because of the cycle-and-add property of the C/A-codes, Eq. (A8) becomes the following:

$$\begin{aligned} I_{j3}(t) &= A_j C_k(t) \cos(2\pi \Delta f_{ij} t + \phi_j - \phi_i) \\ Q_{j3}(t) &= A_j C_k(t) \sin(2\pi \Delta f_{ij} t + \phi_j - \phi_i) \end{aligned} \tag{A9}$$

where $C_k(t)$ is another code in the same family, which is different for each value of N. The signal components at ④ are then as follows:

$$\begin{aligned} I_{i4}(t) &= A_i T \\ Q_{i4}(t) &= 0 \end{aligned} \tag{A10}$$

$$\begin{aligned} I_{j4}(t) &= A_j \int_0^T C_k(t) \cos(2\pi \Delta f_{ij} t + \phi_j - \phi_i)\, \mathrm{d}t \\ Q_{j4}(t) &= A_j \int_0^T C_k(t) \sin(2\pi \Delta f_{ij} t + \phi_j - \phi_i)\, \mathrm{d}t \end{aligned} \tag{A11}$$

where T is a multiple M of 1 ms C/A-code repetition periods.

Power in the two correlated signals is given as follows:

$$2P_{i4} = I_{i4}^2 + Q_{i4}^2 = A_i^2 T^2 \tag{A12}$$

$$\begin{aligned} 2P_j(\Delta_{ij}) &= I_{j4}^2 + Q_{j4}^2 \\ &= A_j^2 \left\{ \left[\int_0^T C_k(t) \cos(2\pi f_{ij} t)\, \mathrm{d}t \right]^2 + \left[\int_0^T C_k(t) \sin(2\pi f_{ij} t)\, \mathrm{d}t \right]^2 \right\} \end{aligned} \tag{A13}$$

Note that, through expansion using trigonometric identities, the dependence upon the phase difference has been removed in Eq. (A13), which resembles the expression for a Fourier power spectrum component at the Doppler difference. The code repeats at a 1-kHz rate, and the integration over each 1 ms code period is identical for Doppler differences of multiples of 1 kHz. Thus, the ratio of Eq. (A13) to Eq. (A12) the cross-correlation power ratio, can be reduced to the following:

$$P_{ij}(n) = \frac{A_j^2}{A_i^2}\left\{\left[\frac{1}{2L}\int_0^{2L} C_k \cos\left(\frac{n\pi}{L}t\right) \mathrm{d}t\right]^2 + \left[\frac{1}{2L}\int_0^{2L} C_k \sin\left(\frac{n\pi}{L}t\right) \mathrm{d}t\right]^2\right\} \tag{A14}$$

where

$$\Delta f_{ij} = n/2L = 1000\,n \text{ Hz}$$

$$2L = 0.001 \text{ s} = T/M$$

The term in the brackets of Eq. (A14) can be recognized as the nth power spectrum component. Obviously, Eq. (A13) can take on values for Doppler differences other than multiples of 1 kHz, depending upon the value of T, but peaks at the 1 kHz lines. That is, there is cross-correlation at other Doppler differences, but less than at the 1 kHz lines. Because of the navigation message databits, M is limited to 20. If T were allowed to be infinite (very large), the cross correlation at these other values would approach zero. For example, if T was a large multiple of Doppler difference cycle periods, integration over each cycle would pick up a different portion of the code, or multiple codes plus a different portion of the code. If we take a Doppler difference of 50 Hz and integrate over 20 ms, then integration over the last 10 ms would cancel that over the first 10 ms, because the codes would simply be flipped. This is true for any Doppler difference that is a submultiple of 1 kHz and an integration time that is an integer number of Doppler difference cycle periods. For other Doppler differences, other than the multiples of 1 kHz, a longer integration period is required for eventual cancellation.

Equation (A14) was evaluated for equal amplitude signals for all relative code phases as defined in Eq. (A6) for two specific GPS C/A-code pairs. The first pair (PRN6/PRN28) is considered to be the *worst* pair of GPS codes; whereas the second pair (PRN7/PRN201) is made up of the *best* GPS code and the *best* selected for Inmarsat-3.[9] For each pair, the maximum spectrum components over all possible integer code phases ($0 \leq N \leq 1022$), are plotted in Fig. A2 for Doppler differences up to 20 kHz. The maximum spectrum components are also plotted in Fig. A3, for Doppler differences in the range 1017–1029 kHz, which corresponds to one of the signals transmitting in the first null of the other. This ± 6 kHz Doppler range represents the maximum expected from satellite and user motion. The average over all N computed for the PRN6/PRN28 pair is also plotted in Fig. A3 for comparison. The average is approximately equal to the following spectral line envelope:

$$S(f) = \frac{1}{1023}\,\frac{\sin^2(n\pi/1023)}{(n\pi/1023)^2} \tag{A15}$$

which is about 8–9 dB below the maximum values. Note that in both figures, it doesn't seem to matter which code pair is used when determining the worst-case

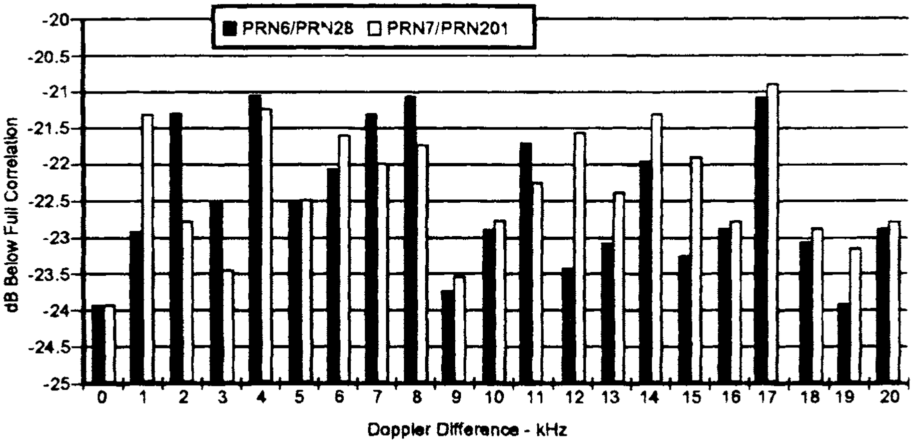


Fig. A2 Maximum cross-correlation spectral components for two code pairs in Doppler difference range of 0–20 kHz.

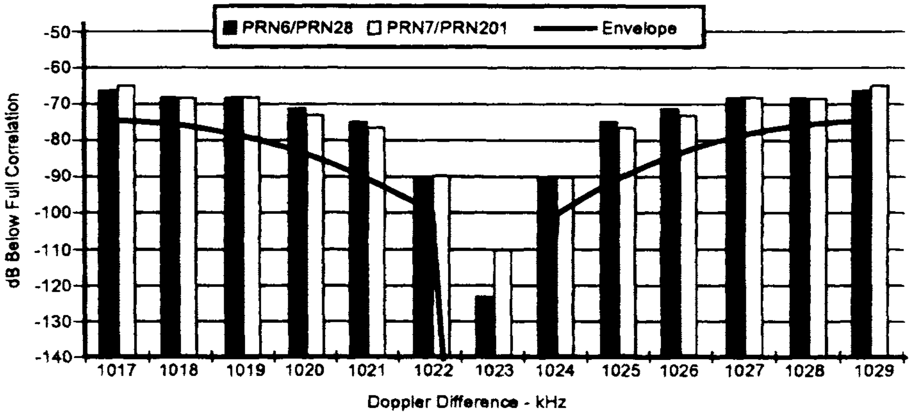


Fig. A3 Maximum cross-correlation spectral components for two code pairs in Doppler difference range of 1017–1029 kHz.

magnitudes. This is because each of the 1023 code phases ($0 \leq N \leq 1022$) results in a different code. Consequently, codes with bad (i.e., large) spectral components are generated in each case, and some of these codes are unbalanced, as well.

Although some of the *worst*-case components shown in Fig. A3 slightly exceed the predicted −80 dB level stated in Sec. IIB1, they are still typically below −70 dB. Within ±2 kHz of the null, they are, in fact, below −80 dB. More significantly, all components are a good 50–60 dB lower than the level near center frequency. This is in addition to the margin realized from pulsing with a 1/11 duty cycle. (See Appendix B)

The derivation presented above applies when the cross-correlation code transitions are lined-up, which, of course, they rarely will be. Spilker[10] points out that, at the *n* kHz carrier frequency differences, there also exists a code frequency shift, which is less by a factor of 1540. Thus, for example, at the 1 kHz carrier frequency difference, there is a code frequency difference of 1000/1540 ≈ 0.65 chips per second. Thus, the code transitions will not stay lined-up. This is especially true for a large frequency difference of 1.023 MHz, in which case the code frequency difference is approximately 664 chips per second. McGraw[12] pointed out that cross-correlation levels can be even higher when the code transitions are not lined up. However, because of the rapidly changing time relationships between the codes, the resulting cross-correlation becomes noise-like, and simply becomes a noise interference. The effect of this interference is addressed in Appendix B.

Appendix B: Interference Caused by Pseudolite Signal Level

One C/A-code signal can interfere with another if it is strong enough, independent of the cross-correlation. This can certainly happen in the case of a PL signal, which may interfere with satellite signals as well as with other PL signals. In other words, a PL signal is a noise source that may jam the other signals unless measures are taken to mitigate this effect. The following provides an assessment of the impact of that jamming and a method for minimizing it; namely, by pulsing the PL signal.

Background

In general, the spreading process in a receiver's correlator is defined as a signal (noise or otherwise) being passed through a filter described with a frequency response equal to the spectral density of the PRN code. That is, the interference noise density at the output of the correlator is as follows:

$$N_{0I} = \int_{-\infty}^{\infty} S_c(f) S_I(f)\, \mathrm{d}f \tag{B1}$$

where $S_c(f)$ is the spectral density of the reference PRN code and $S_I(f)$ is the density of the interference or noise. The reference C/A-code has a discrete spectral density that can be described as follows:

$$S_c(f) = \sum_{n=-\infty}^{\infty} c_n \delta(f - 1000\, n) \tag{B2}$$

where the c_n are spectral line coefficients: $\delta(f)$ is the dirac delta function; and the c_n vary about the envelope of Eq. (A15) in Appendix A. The reference C/A-code spectral density has the property that

$$\int_{-\infty}^{\infty} S_c(f)\, \mathrm{d}f = 1 \tag{B3}$$

First consider bandlimited thermal noise with density N_0 with a two-sided intermediate frequency (IF) bandwidth of B_I (brick wall filter).

$$N_{0T} = N_0 \int_{-B_I/2}^{B_I/2} S_c(f)\, \mathrm{d}f \leq N_0 \tag{B4}$$

The variation of the C/A-code spectral lines averages out over the wide bandwidth.

A similar equation applies for wide- and narrow-band interference with a spectral density as follows:

$$S_I(f) = \frac{P_I}{f_u - f_l} \tag{B5}$$

for upper and lower frequency limits f_u and f_l (converted to IF) and total interference power P_I (relative to the signal power S).

$$N_{0I} = \frac{P_I}{f_u - f_l} \int_{\max(-\alpha_1, \alpha_3)}^{\min(\alpha_1, \alpha_2)} S_c(f)\, \mathrm{d}f \tag{B6}$$

where

$$\alpha_1 = B_I/2,\ \alpha_2 = f_u - f_{IF} \quad \text{and} \quad \alpha_3 = f_l - f_{IF}.$$

For narrow bandwidth noise interference centered at the GPS frequencies (i.e., somewhat less than the code chipping rate $1/T_c$), this equation becomes as follows:

$$N_{0I} \approx P_I T_c \tag{B7}$$

This is only true for the C/A code where interference bandwidths are on the order of 100 kHz or greater because of the variations in the line spectrum of the codes.

Pseudolite Interference

Now assume that the interference is another C/A-code signal with the following spectral density:

$$S_{\mathrm{PL}}(f) = P_{\mathrm{PL}} \sum_{n=-N_B}^{N_B} c_n' \delta(f - 1000\, n \pm 1.023 \times 10^6) \tag{B8}$$

where P_{PL} is the received PL signal power, and N_B indicates the band-limiting effect. Because the summations in Eqs. (B2) and (B8) are over a large number of spectral components, it suffices to use the average envelope of Eq. (A15), divided by 1000 to spread the components into a continuous spectral density, for evaluation. Then, Eq. (B1) can be approximated with the following integral:

$$N_{0\mathrm{PL}} \approx P_{\mathrm{PL}} T_c^2 \int_{-B_I/2}^{B_I/2} \frac{\sin^2(\pi f T_c) \sin^2[\pi(f \pm 1/T_c)T_c]}{(\pi f T_c)^2 [\pi(f \pm 1/T_c)T_c]^2}\, \mathrm{d}f \tag{B9}$$

This can be compared to the interference noise density from one GPS satellite j to another satellite i at the normal frequency given by the following:

$$N_{0S_{ij}} \approx P_{S_j} T_c^2 \int_{-B_I/2}^{B_I/2} \frac{\sin^4(\pi f T_c)}{(\pi f T_c)^4}\, \mathrm{d}f \tag{B10}$$

In general, Eqs. (B9) and (B10) must be evaluated numerically. This was done over a wide bandwidth of $\pm 10/T_c$, resulting in the following relationship:

$$10 \log_{10}\left(\frac{N_{0\mathrm{PL}}}{N_{0S_{ij}}}\right) = 10 \log_{10}\left(\frac{P_{\mathrm{PL}}}{P_{S_j}}\right) - 8.19\ \mathrm{dB} \tag{B11}$$

which indicates that the PL interferes by 8.19 dB less by transmitting in the null, than it would if it were transmitting at the same frequency as the satellites. However, this is not enough!

The first term of Eq. (B11) can become significant as the user receiver comes closer to a PL. This is a key reason for PL pulse modulation, which has the effect of only interfering a percentage of the time equal to the pulse duty cycle, provided that the receiver clips the pulses. This results in a loss in received satellite C/N_0 of either:

$$L = -10 \log_{10}\left[1 - DC + \frac{2R_{\Delta f}}{3}\left(\frac{L_{\max}}{L_N}\right)^2 \frac{DC}{(1 - DC)^2} B_I T_c\right] \qquad \text{(B12)}$$

or

$$L = 10 \log_{10}\left[1 + \frac{2R_{\Delta f}}{3}\frac{P_{PL}}{N_0} T_c \cdot DC\right] \qquad \text{(B13)}$$

depending upon whether or not the PL signal is saturating the receiver's analog-to-digital (A/D) sampler (*soft* or *hard-limiting*), where $R_{\Delta f}$ is the reduction in interference realized using a frequency offset (0.1517, if ±1.023 MHz, and 1, if on frequency), $L_{\max}$ is the maximum sampler threshold level, L_N is the one sigma noise level in terms of the threshold level and P_{PL}/N_0 is the average received PL signal-to-noise density. Eq. (B12) represents the loss when the PL pulses are saturating the A/D and Eq. (B13) represents the loss when they are not. Note that in the former case, the loss is proportional to the ratio of IF bandwidth to the code bandwith.

Cross-correlation can still occur, even if the pulses are clipped. However, as was shown in Appendix A, the cross-correlation is reduced substantially by PL transmission in the first null of the satellite C/A-code/spectrum (L_1 ± 1.023 MHz). Pseudolite transmission at higher nulls is also possible, but this begins to add more complexity to a receiver designed to process both GPS and PL signals. Another key reason for pulsing is to prevent capturing the front end of the user's receiver. This is especially important for hard-limiting receivers (1-bit samplers) that normally do not employ front-end AGC circuits.

Appendix C: Navigation Filter Modeling with Pseudolite Measurements

A GPS pseudorange measurement (and carrier range) is generally a nonlinear function of the satellite position vector for ($\boldsymbol{r}_s$) and the user position vector ($\boldsymbol{r}_u$). This measurement is modeled by the following:

$$z = g + b_u + \nu \qquad \text{(C1)}$$

where

$$g = |\boldsymbol{r}_s - \boldsymbol{r}_u| \qquad \text{(C2)}$$

is the geometric range; b_u is the user clock offset from GPS time, and v represents the composite of uncorrected measurement errors caused by atmospherics, satellite timing offsets, multipath, and noise. Given the satellite ephemeris and an a priori estimate of the user location ($\boldsymbol{r}_\mathrm{u}$), then g can be represented by the Taylor series expansion:

$$g = g_0 + \boldsymbol{h}\Delta\boldsymbol{r}_\mathrm{u} + \Delta\boldsymbol{r}_u^T J \Delta\boldsymbol{r}_\mathrm{u}/2 + \cdots \tag{C3}$$

where

$$\Delta\boldsymbol{r}_\mathrm{u} = \boldsymbol{r}_\mathrm{u} - \boldsymbol{r}_\mathrm{u}^- \tag{C4}$$

$$g_0 = |\boldsymbol{r}_\mathrm{s} - \boldsymbol{r}_\mathrm{u}^-| \tag{C5}$$

$$\boldsymbol{h} = \partial g/\partial\boldsymbol{r}_\mathrm{u}|_{\boldsymbol{r}_\mathrm{u} = \boldsymbol{r}_\mathrm{u}^-} \tag{C6}$$

$$J = \partial^2 g/\partial\boldsymbol{r}_\mathrm{u}\partial\boldsymbol{r}_\mathrm{u}|_{\boldsymbol{r}_\mathrm{u} = \boldsymbol{r}_\mathrm{u}^-} \tag{C7}$$

Linear Measurement Model

Given the large user/satellite separation and slowly changing geometry, common practice is to employ an extended (linearized) Kalman filter that encompasses only the first two terms for g in Eq. (C3). The standard EKF equations for updating the a priori estimate of user position ($\boldsymbol{r}_u$) and clock bias (b_u) are as follows:

$$\begin{aligned} \boldsymbol{x}_\mathrm{u}^+ &= \boldsymbol{x}_\mathrm{u}^- + \boldsymbol{k}(z - g_0) \\ \boldsymbol{k} &= P^-\boldsymbol{h}^T/(\boldsymbol{h}P^-\boldsymbol{h}^T + \sigma_v^2) \\ P^+ &= (I - \boldsymbol{kh})P^-(I - \boldsymbol{kh})^T + \boldsymbol{k}(\sigma_v^2)\boldsymbol{k}^T \end{aligned} \tag{C8}$$

where

$$\boldsymbol{x}_\mathrm{u} \equiv \begin{pmatrix} \boldsymbol{r}_\mathrm{u} \\ \text{---} \\ b_u \end{pmatrix} \tag{C9}$$

and

$$\boldsymbol{h} \equiv [(\boldsymbol{r}_s - \boldsymbol{r}_u)^{T/g} \quad 1] \tag{C10}$$

Also, P^- is the covariance of the a priori state estimate ($\boldsymbol{x}_\mathrm{u}$), and σ_v^2 is the variance of the measurement error v, assumed to be Gaussian white noise.

Nonlinear Measurement Model

When PL measurements are introduced, the user/PL range is much less and more dynamic. Consequently, a Gaussian second-order filter[19] can be employed that accounts for the measurement nonlinearity with the quadratic component included

$$z - g_0 = \boldsymbol{h}\Delta\boldsymbol{x}_\mathrm{u} + \Delta\boldsymbol{r}_\mathrm{u}^T J \Delta\boldsymbol{r}_\mathrm{u}/2 + v \tag{C11}$$

The corresponding GSO filter equations are expressed by the following:[20]

$$
\begin{aligned}
\boldsymbol{x}_0^+ &= \boldsymbol{x}_u^- + \boldsymbol{k}(z - g_0 - \eta) \\
\boldsymbol{k} &= P^- \boldsymbol{h}^T / (\boldsymbol{h} P^- \boldsymbol{h}^T + \sigma_v^2 + \sigma_n^2) \\
P^+ &= (I - \boldsymbol{k}\boldsymbol{h}) P^- (I - \boldsymbol{k}\boldsymbol{h})^T + \boldsymbol{k}(\sigma_v^2 + \sigma_\eta^2)\boldsymbol{k}^T \\
\eta &= \mathrm{Tr}[JP^-]/2 \\
\sigma_\eta^2 &= \mathrm{Tr}[JP^- JP^-]/2
\end{aligned}
\qquad \text{(C12)}
$$

where $\boldsymbol{x}_u$, $\boldsymbol{h}$, and J are as defined in this appendix. The new components account for the bias and measurement variance introduced by the quadratic nonlinearity.

References

[1] Van Dierendonck, A. J., Elrod, B. D., and Melton, W. C., "Improving the Integrity, Availability and Accuracy of GPS Using Pseudolites," *Proceedings of NAV '89* (London, UK), Royal Institute of Navigation, London, UK, Oct. 17–19, 1989, (RION Paper 32).

[2] Schuchman, L., Elrod, B. D., and Van Dierendonck, A. J., "Applicability of an Augmented GPS for Navigation in the National Airspace System," *Proceedings of IEEE,* Vol. 77, No. 11, Nov. 1989, pp. 1709–1727.

[3] Cohen, C. A., et al., "Real-Time Cycle Ambiguity Resolution using a Pseudolite for Precision Landing of Aircraft Using GPS," *Proceedings of DSNS '93,* Amsterdam, The Netherlands, March 31–April 3, 1993.

[4] Harrington, R. L., and Dolloff, J. T., "The Inverted Range: GPS User Test Facility," *Proceedings, IEEE Position, Location, and Navigation Symposium,* (PLANS '76, San Diego, CA), IEEE, New York, Nov. 1976, pp. 204–211.

[5] Klein, D., and Parkinson, B. W., "The Use of Pseudo-Satellites for Improving GPS Performance," *Global Positioning System,* Vol. II, Institute of Navigation, 1986, pp. 135–146.

[6] Stansel, T. A., Jr., "RTCM SC-104 Recommended Pseudolite Signal Specification," *Global Positioning System,* Vol. III, Institute of Navigation, 1986, pp. 117–134.

[7] Van Dierendonck, A. J., "The Role of Pseudolites in the Implementation of Differential GPS," *Proceedings, IEEE Position, Location, and Navigation Symposium,* (PLANS '90, Las Vegas, NV), IEEE, New York, March 1990.

[8] Elrod, B. D., and Van Dierendonck, A. J., "Testing and Evaluation of GPS Augmented with Pseudolites for Precision Landing Applications," *Proceedings of DSNS'93,* Amsterdam, The Netherlands, March 31, 1993.

[9] Nagle, J., Van Dierendonck, A. J., and Hua, Q. D., "Inmarsat-3 Navigation Signal C/A Code Selection and Interference Analysis," *Navigation,* Winter 1992–93, pp. 445–461.

[10] Spilker, J. J., Jr., "GPS Signal Structure and Performance Characteristics," *Global Positioning System,* Vol. I, Institute of Navigation, 1980, pp. 29–54.

[11] "Global Orbiting Navigation Satellite System (GLONASS) Interface Control Document," RTCA Paper 518-91/SC159-317 (Handout—Jan. 1992).

[12]McGraw, G. A., "Analysis of Pseudolite Code Interference Effects for Aircraft Precision Approaches," *Proceedings of ION 50th Annual Meeting* (Colorado Springs, CO), Institute of Navigation, Washington, DC, June 6–8 1994.

[13]Green, G. B., et al., "The GPS 21 Primary Satellite Constellation," *Navigation,* Spring 1989, pp. 9–24.

[14]Anon, "1992 Federal Radionavigation Plan." U.S. Dept. of Transportation, Dec. 1992, pp. 2–15.

[15]Anon., "Recommendations of the Carrier Phase Working Group to RTCM SC-104," RTCM Paper 170-921/SC104-92, Aug. 10, 1992.

[16]Anon., "Wide Area Augmentation System Signal Specification," App. 2 in U.S. DOT/FAA Specification for the Wide Area Augmentation System (WAAS), FAA-E-2892, May 9, 1994.

[17]Anon., "Minimum Aviation System Performance Standards—DGNSS Instrument Approach System: Special Category I (SCAT-I)," RTCA/DO-217, RTCA, Inc., Washington, DC, Aug. 27, 1993.

[18]Elrod, B. D., Barltrop, K. J., and Van Dierendonck, A. J., "Testing of GPS Augmented with Pseudolites for Precision Approach Applications," Proceedings of ION GPS '94 (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–23, 1994.

[19]Jazwinski, A. H., *Stochastic Processes and Filtering Theory,* Academic Press, New York, 1970, pp. 340–346.

[20]Widnall, W. S., "Enlarging the Region of Convergence of Kalman Filters Employing Range Measurements," *AIAA Journal,* Vol. 2, No. 3, 1973, pp. 283–287.

Chapter 3

Wide Area Differential GPS

Changdon Kee*
Stanford University, Stanford, California 94305

I. Introduction

IN addition to reducing cost and complexity, the GPS is expected to improve the accuracy of navigation greatly for land, marine, and aircraft users. Under normal operating conditions, it can provide positioning accuracies in the range of 15–25 m. However, with selective availability (SA) the errors incurred by typical civilian users have been found to be 100 m or more. In some situations, especially for precision landing of an aircraft or for harbor navigation, these accuracies are insufficient.

Differential GPS (DGPS) is a means for improving navigation accuracy in a local area. A single DGPS monitor station at a known location can compute a range error correction for each GPS satellite in view. These error corrections are then broadcast to users in the vicinity, as depicted in Fig. 1. By applying the corrections to the signals received, a user can typically improve the accuracy down to the 2–5 m level (see Chapter 1 of this volume and Refs. 1 and 2). However, as the distance between the user and the monitor station increases, range decorrelation occurs, and accuracy degrades. This increased error is caused by ephemeris error, ionospheric time delay error, and tropospheric error. As the user and reference station separate, the projection of the ephemeris error onto the user–satellite line of sight is no longer the same as that projected onto the monitor station–satellite line of sight, as illustrated in Fig. 2. (The accuracy of the range correction broadcast by a DGPS monitor station degrades with distance. The figure shows an ephemeris error δR that produces a small range error δR_s at the monitor station but a larger range error δR_u at the user location. If the user were to employ the range correction broadcast by the monitor station, a residual range error of $\delta R_s - \delta R_u$ would remain.)

*WADGPS Algorithm Development Group Leader, WADGPS Laboratory, Department of Aeronautics and Astronautics, HEPL (GP-B).

The maximum range error difference δR_{error} between the monitor and user is given by the following:

$$\max(\delta R_{\text{error}}) \approx \frac{d}{D}\,\delta R \tag{1}$$

where δR is the magnitude of the satellite ephemeris error; d is the separation between the user and the monitor station; and D is the distance from the user to the GPS satellite.

If the two receivers are widely separated, the lines of sight through the ionosphere are also different, resulting in differences in the ionospheric delay observed. A similar, but smaller effect occurs for the tropospheric delay.

Beyond a separation distance of 100 km, a scalar range error correction is not sufficiently accurate to realize the full potential of DGPS. In fact, hundreds of monitor stations would be required to provide standard single-station DGPS aiding across the entire United States. Wide area differential GPS (WADGPS)[3] provides a powerful means for bridging the gap between unaided performance and high-accuracy navigation in the vicinity of a correction station. Now the Federal Aviation Administration (FAA) is planning to implement WADGPS in the National Airspace System by 1997. Various WADGPS techniques have been suggested by Refs. 4–8.

The following sections describe the WADGPS architecture, and master station algorithms. Then user message content and format, and WADGPS error budget are discussed. The next sections describe WADGPS simulations and evaluations using actual field data. The next chapter in this volume discusses implementation of WADGPS for the National Airspace System.

II. Wide Area Differential GPS Architecture and Categories

Instead of calculating a scalar range error correction for each satellite, as is done in DGPS, WADGPS provides a vector of error corrections composed of a three-dimensional ephemeris error and clock offset for each GPS satellite, plus ionospheric time delay parameters. The accuracy of the WADGPS correction is nearly constant within the monitored region, and degrades gracefully on the perimeter.

A. Wide Area Differential GPS Architecture

The WADGPS network includes at least one master station, a number of monitor stations, and communication links. Each monitor station is equipped with a high-quality clock and a high-quality GPS receiver capable of tracking all satellites within the field of view. The GPS measurements are taken at each monitor station and sent to the master station. The master station computes GPS error components, based on the known monitor station locations and the information collected. The computed error corrections are transmitted to the users via any convenient communication link, such as satellite, telephone, or radio.

Figure 3 provides an overview of the WADGPS, and Fig. 4 shows the flow of information between the system components. The process can be summarized as follows:

1) Monitor stations at known locations collect GPS pseudoranges from all satellites in view.

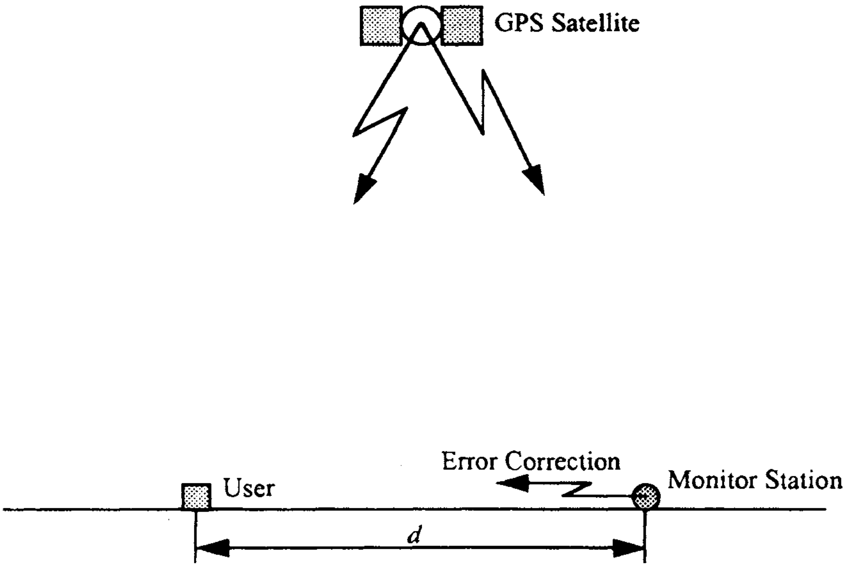


Fig. 1 Overview of differential GPS.

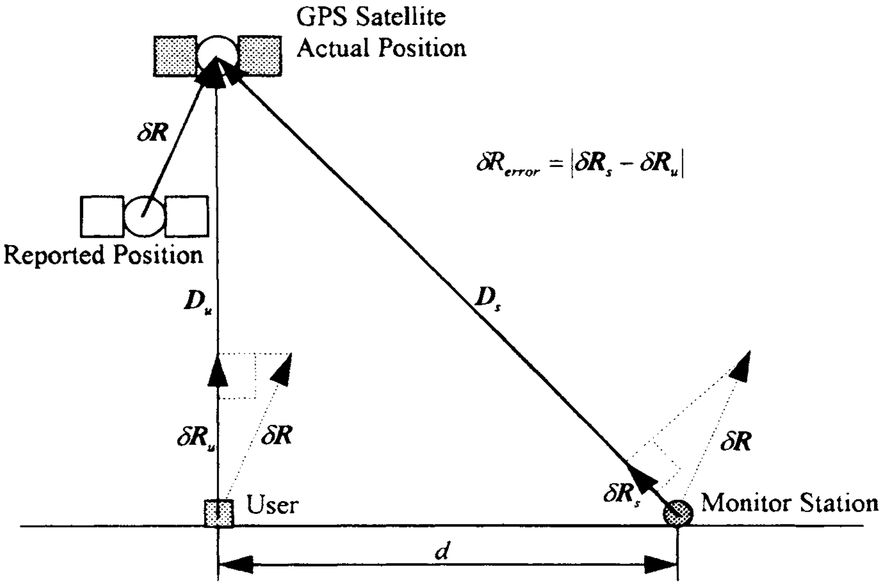


Fig. 2 Degradation of DGPS accuracy with distance.

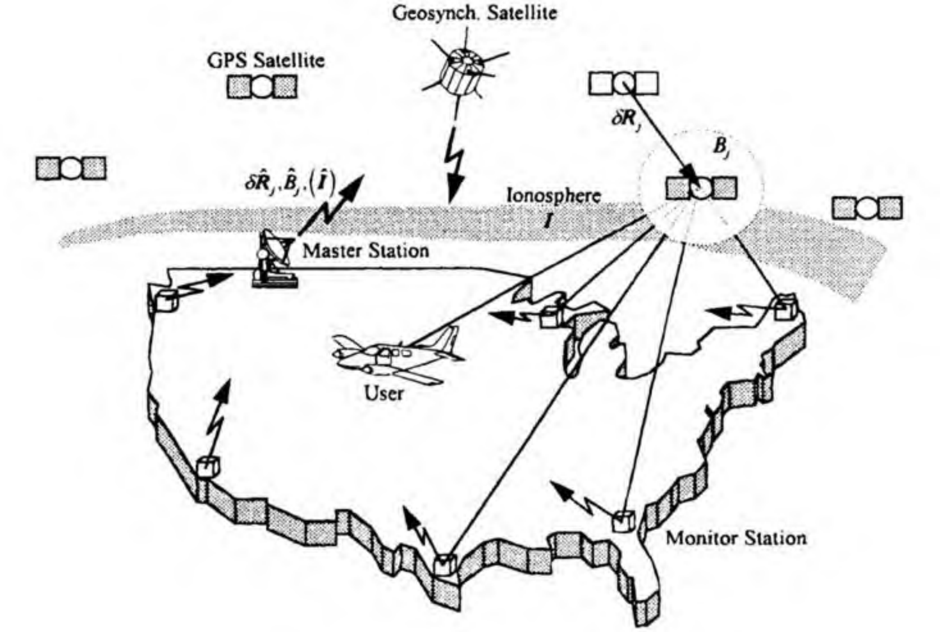


Fig. 3 Wide area differential GPS concept.

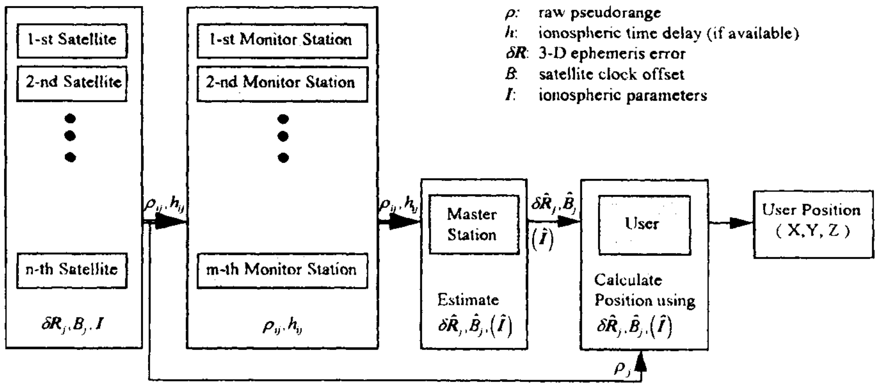


Fig. 4 Block diagram of WADGPS components.

2) Pseudoranges and dual-frequency ionospheric delay measurements (if available) are sent to the master station.

3) Master station computes an error correction vector.

4) Error correction vector is transmitted to users.

5) Users apply error corrections to their measured pseudoranges and collected ephemeris data to improve navigation accuracy.

B. Wide Area Differential GPS Categories

The WADGPS system can be categorized by the estimator located at the master station. In the design of WADGPS, we must address such issues as the receiver required for monitor station and user, the estimation speed, which corresponds to the update rate of error corrections, and the navigation accuracy. Because the most important application of WADGPS is aviation, the navigation accuracy should be the major concern.

The master station estimates the three-dimensional ephemeris errors, the satellite clock errors, the monitor station receiver clock errors, and, optionally, the ionospheric time delay parameters. It does not transmit the monitor station receiver clock errors.

A single-frequency receiver normally provides L_1 pseudorange and continuous carrier phase as outputs. A dual-frequency receiver provides not only L_1 pseudorange and continuous carrier phase but also L_2 pseudorange and continuous carrier phase from which the ionospheric time delay can be calculated as if it were an extra output, but dual-frequency receivers are far more expensive than single-frequency receivers.

A single-frequency receiver can, in principle, estimate the ionospheric time delay by measuring the dispersive effect of the ionospheric time delay on the received code and carrier. Using the fact that the ionospheric time delays in the pseudorange and continuous carrier phase are equal in size and have opposite signs, we may be able to estimate ionospheric time delay with a single-frequency receiver[9,10] at the expense of a loss in navigation accuracy. However, the single-frequency technique needs further study. Because dual-frequency receivers (provided they are at both the monitor station and the user) directly give the ionospheric time delays, they can save time that would have to be spent on estimating the ionospheric time delay parameters in the master station and can, therefore, improve the navigation accuracy.

Three WADGPS algorithms (A, B, and C) and their performances are summarized in Table 1 and Table 2, respectively, and each algorithm is discussed in the following subsections.

1. Algorithm A

This algorithm allows both monitor station and user to use single-frequency receivers that do not provide ionospheric time delays as extra measurements. In this algorithm, the master station estimates the three-dimensional ephemeris errors, the satellite clock errors, and the ionospheric time delay parameters in one large filter using pseudoranges as the only measurement vector. Processing

Table 1 Wide area differential GPS algorithms

	Algorithms		
	A	B	C
Variables to be estimated			
Three-dimensional ephemeris errors	Yes	Yes	Yes
Satellite clock error + SA	Yes	Yes	Yes
Ionospheric parameters	Yes	Yes	No
Variables to be transmitted (error corrections)			
Three-dimensional ephemeris errors	Yes	Yes	Yes
Satellite clock error + SA	Yes	Yes	Yes
Ionospheric parameters	Yes	Yes	No
Number of master station estimators	One	Two	One
Size of master station estimator	Large	Small	Small
Required receiver			
Monitor station	Single-frequency	Dual-frequency[a]	Dual-frequency[a]
User	Single-frequency	Single-frequency	Single-frequency[b] Dual-frequency[a]

[a] There are several receivers, such as the Trimble 4000SSE, and Allan Osborne, Rogue, that can measure the ionospheric time delay even when the P-code is encrypted.

[b] Ionospheric time delay estimation with a single-frequency receiver has been demonstrated by Refs. 9 and 10.

time is the longest among three algorithms because the observation matrix is large. The transmission message consists of the ephemeris errors, the satellite clock errors, and the ionospheric parameters. There are some advantages in terms of lower cost to using this algorithm, the penalty being higher computational load and worse accuracy.

2. *Algorithm B*

By using the extra measurement of ionospheric time delay from a dual-frequency receiver in the monitor station we can separate the one large estimator used in Algorithm A into two small estimators. The estimation of ionospheric time delay parameters is one process, and the estimation of the three-dimensional ephemeris errors, the satellite clock errors, and the monitor station receiver clock errors is another, separate, process.[3]. For the resulting algorithm (algorithm B), a dual-frequency receiver is required in the monitor station, but the user needs only a single-frequency receiver. Because the ionospheric time delays are separate measurements in a dual-frequency receiver, they are used to estimate the ionospheric parameters directly, and the pseudoranges, corrected for ionospheric time delay and tropospheric error, are fed into the other filter, which estimates ephemeris errors and clock errors.

Table 2 Performances of wide area differential GPS algorithms

	Algorithms		
Performance	A	B	C
Computational load	Heavy	Light	Lightest
Navigation accuracy	Good	Better	Best

Relative to algorithm A, algorithm B has better accuracy and reduced computational load. These advantages result from the extra ionospheric measurements and the estimator is divided into two small ones, resulting in reduced matrix sizes. The transmission message to users consists of three-dimensional ephemeris errors and satellite clock errors, as well as ionospheric parameters. In principle, single- or dual-frequency users can take advantage of these error corrections to improve positioning accuracy.

3. Algorithm C

If the mobile users can measure the ionospheric time delay, the ionospheric parameters do not need to be estimated in the master station, and as a result, only three-dimensional ephemeris errors and satellite clock errors need to be estimated there. The resulting algorithm (algorithm C) requires a dual-frequency receiver in the monitor station. Because there is no ionospheric parameter estimation, the transmission message does not contain ionospheric parameters, and therefore, the computational load is smaller than that of algorithm B. A user may be equipped with a dual-frequency receiver, but such a receiver may not be required. We may be able to estimate ionospheric time delay with a single-frequency receiver at the expense of a loss in navigation accuracy.[9] The single-frequency technique needs further study, however.

If the user estimate of ionospheric time delay is accurate, then algorithm C is the most accurate. It is most accurate because it does not fit the ionosphere to the model, as is done in algorithm B. This is especially important in the equatorial and polar regions. Also T_{gd} may be an error source for algorithm B. T_{gd} is a time delay between L_1 and L_2 frequencies in the GPS satellite and is included in the ionospheric time delay measurement from dual-frequency receiver unless it is carefully calibrated and taken off the raw ionospheric time delay measurement. Thus, of the three algorithms discussed, algorithm C provides the best accuracy for users with dual-frequency receivers, but users may opt for a single-frequency receiver depending upon how much accuracy they desire.

C. User Message Content and Format

Transmission of the WADGPS correction could be accomplished by any of the following: geosynchronous satellite broadcast (see next chapter), FM subcarrier,[11] or any other suitable broadcast system. The correction could be converted to the standard differential message format developed by the Radio Technical Commission for Maritime Service Special Committee 104 (RTCM 104).[12] This

Table 3 Wide area differential GPS correction message content

Message		Update rate
SV ID Time tag (GPS time)	PRN Time of transmission	Every 5–10 s
SV clock offset	Offset Offset rate	Every 5–10 s
SV position error (in WGS-84 frame)	X component Y component Z component	Every 1–5 min
Ionospheric parameters	Eight parameters	Every 2–5 min

allows use of receivers designed to meet the current DGPS industry standard without significant modifications.

Clock offsets including SA[13] have been observed to have variation with time constants on the order of three minutes. Thus, an update rate of 0.1–0.2 Hz is sufficient to eliminate the clock error, assuming that users compute WADGPS error correction rate based on prior correction message and apply the correction rate to calculate the present error correction.

Ephemeris errors have been observed to have variations with time constants on the order of 0.5–6 h. Thus, an update rate of 1–5 min is sufficient to eliminate the ephemeris errors.

Usually the total electronic content at zenith varies very slowly (on the order of 6–12 h), but the scintillation of the ionosphere, an abrupt change of the ionosphere in a small region, can make it difficult to estimate. Space vehicle identification and a time tag in GPS time are attached to the beginning of each message. Suggested message content for transmission of the WADGPS correction is shown in Table 3. A more detailed message format is given in the next chapter.

D. Error Budget

The navigation accuracy that a user can achieve using WADGPS is summarized in Table 4. Selective availability is included in the satellite clock offset because part of it is generated by satellite clock dithering. Receiver noise can be decreased by averaging 10 measurements in time. The multipath effect can be reduced by smoothing the code with continuous carrier phase information. This can be achieved with a Hatch/Eshenbach or Kalman filter.

III. Master Station Error Modeling

The key to WADGPS is the formulation and computation of the error correction vector by the master station. This correction consists of a three-dimensional ephemeris error and clock bias for each GPS satellite in view of one or more of the monitor stations, plus eight ionospheric time delay parameters. These parameters are estimated based upon the information gathered by the monitor stations. In addition to the error correction vector, the master station must also estimate the offset of each monitor station clock from a single reference.

The following subsections describe the sources of error, the models used by the master station, and the techniques for estimating the model parameters. For algorithms A and B, the master station computes the correction vector, which is ephemeris and clock errors, and ionospheric time delay parameters, in a 2-step process. In the first step, the parameters in the ionospheric model are identified by a nonlinear static estimation (NSE) algorithm or a recursive filter. Also, there is an alternative ionospheric time delay estimation algorithm, which uses modified interpolation technique.[14] The estimated ionospheric delays are then used to adjust the raw measurements from each of the stations. The second stage solves for the ephemeris and clock errors for each of the GPS satellites observed by the network, using a batch least squares (BLS) solution or recursive filter.

A. Ionospheric Time Delay Model for Algorithms A or B

As GPS satellite signals traverse the ionosphere, they are delayed by an amount proportional to the number of free ions encountered (total electron content). The ion density is a function of local time, magnetic latitude, sunspot cycle, and other factors. Its peak occurs at 2:00 p.m. local time.

Klobuchar developed a simple analytical model for ionospheric time delay, which we have used as the basis for the WADGPS ionospheric correction model.[15] His model yields an ionospheric time delay prediction that reduces the rms error by at least 60% for the entire northern hemisphere.[16] We can improve this accuracy by performing a parameter fit optimized for the region of interest.

In Klobuchar's model, the vertical ionospheric time delay is expressed by the positive portion of a cosine wave plus a constant night-time bias, as follows[15]:

$$T_{ij} = A_1 + A_2\cos\,[2\Pi(\tau - A_3)/A_4] \tag{2}$$

where T_{ij} = ionospheric time delay in vertical direction at the intersection of the ionosphere with the line from the ith station to the jth satellite; $A_1 = 5 \times 10^{-9}$ seconds (night-time value); $A_2 = \alpha_1 + \alpha_2\phi_M + \alpha_3\phi_M^2 + \alpha_4\phi_M^3$ (amplitude); A_3 = 14:00 local time (phase); $A_4 = \beta_1 + \beta_2\phi_M + \beta_3\phi_M^2 + \beta_4\phi_M^3$ (period); ϕ_M = geomagnetic latitude of ionosphere subpoint; α_i, β_i = ionospheric parameters (***I***) transmitted by the GPS satellites or by the master station; and, τ = local time.

Table 4 Wide area differential GPS error budget

Source	Error budget, m
Ephemeris errors	0.4
Satellite clock offset/selective availability	0.2
Ionospheric time delay	0.5
Tropospheric error	0.3
Receiver noise[a]	0.2
Multipath effect[b]	0.1
UERE, rms	0.77
Navigation accuracy, rms (HDOP = 1.5)	1.2

[a] Receiver noise is based on averaging 10 measurements.

[b] Multipath effect can be reduced by smoothing the code with continuous carrier phase information.

A typical vertical time delay profile generated by this model is shown in Fig. 5. The delay shown corresponds to an L_1 signal coming from a satellite directly above the observer. To estimate the actual ionospheric time delay h_{ij}, for a given satellite elevation angle, we must scale T_{ij} by the appropriate obliquity factor Q_{ij}, which is defined as the secant of the zenith angle at the mean ionospheric height, as follows:

$$h_{ij} = T_{ij}(\boldsymbol{I}) \cdot Q_{ij}(\theta) \tag{3}$$

where h_{ij} = ionospheric time delay from ith station to jth satellite; $Q_{ij} = {}^{1}/\sin[\sin^{-1}\{r_e/(r_e + h_{\text{iono}}) \cos\theta\}]$ = obliquity factor from ith station to jth satellite; r_e = radius of the Earth; h_{iono} = height of the average ionosphere; θ = elevation angle; and, $\boldsymbol{I} = [\alpha_1, \ldots, \alpha_4, \beta_1, \ldots, \beta_4]^T$ = ionospheric parameter.

The τ and ϕ_M of Eq. (2) are constant at each time-step.

The task of the master station is to generate the eight parameters, $[\alpha_1, \ldots, \alpha_4, \beta_1, \ldots, \beta_4]$, which, when substituted in the Klobuchar model, will yield the best ionospheric delay estimate for the region covered by the WADGPS network.

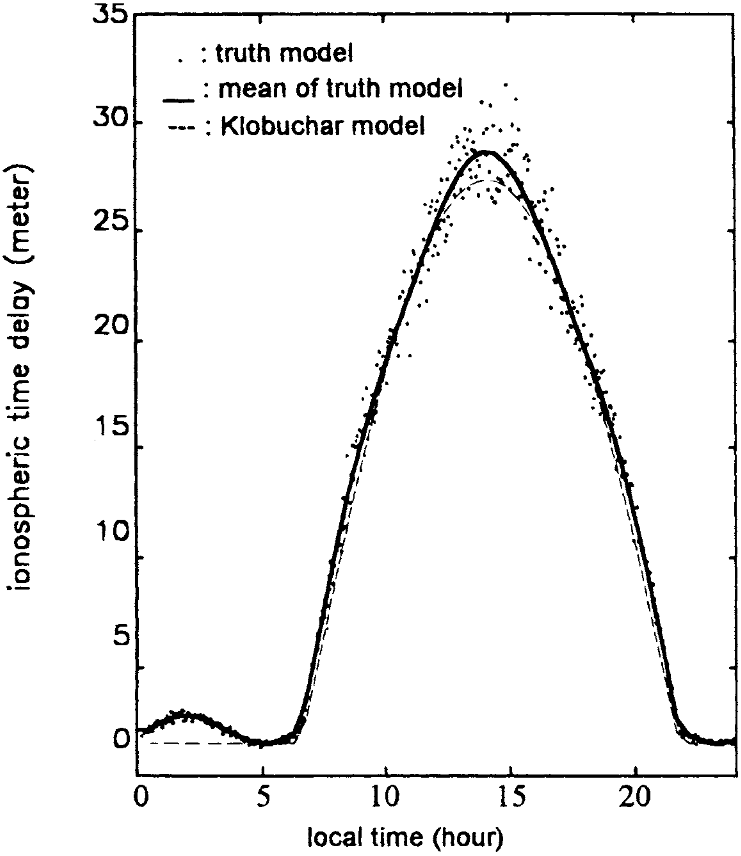


Fig. 5 Klobuchar model (cosine curve) and truth model of ionospheric time delay (at Stanford, California, elevation angle = 90 deg).

1. *Ionospheric Time Delay Measurement Equation*

By collecting h_{ij} for satellite $i = 1, \ldots, m$, and station $j = 1, \ldots, n$, the following ionospheric time delay measurement equation can be obtained:

$$\boldsymbol{d} = \boldsymbol{h}(\boldsymbol{I}) + \boldsymbol{v} \tag{4}$$

where $\boldsymbol{d} = [d_{11}, \ldots, d_{ln} \vdots, \ldots, \vdots d_{ml}, \ldots, d_{mn}]^T$; d_{ij} = ionospheric time delay from *i*th station to *j*th satellite measured using dual-frequency technique; $\boldsymbol{h}(\boldsymbol{I}) = [h_{ll}(\boldsymbol{I}), \ldots, h_{ln}(\boldsymbol{I}) \vdots, \ldots, \vdots h_{ml}(\boldsymbol{I}), \ldots, h_{mn}(\boldsymbol{I})]^T$; and $\boldsymbol{v}$ = measurement noise.

The linearized form of Eq. (4) is as follows:

$$\boldsymbol{\delta d} = \boldsymbol{H} \cdot \boldsymbol{\delta I} + \boldsymbol{v}$$

where

$$\boldsymbol{I} = \boldsymbol{I}_0 + \boldsymbol{\delta I}$$

$$\boldsymbol{I}_0 = [\alpha_{1_0}, \ldots, \alpha_{4_0}, \beta_{1_0}, \ldots, \beta_{4_0}]^T = \text{ nominal ionospheric parameters}$$

$$\boldsymbol{\delta I} = [\delta\alpha_1, \ldots, \delta\alpha_4, \delta\alpha_4, \delta\beta_1, \ldots, \delta\beta_4]^T = \text{ increment of ionospheric parameters}$$

$$\boldsymbol{\delta d} = \boldsymbol{d} - \boldsymbol{h}(\boldsymbol{I}_0)$$

$$\boldsymbol{H} = \left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{I}}\right|_{I=I_0} = [\boldsymbol{T}_{11}^T \cdot Q_{11}, \ldots, T_{ln}^T \cdot Q_{ln} \vdots, \ldots, \vdots \boldsymbol{T}_{ml}^T \cdot \mathrm{Q}_{ml}, \ldots, \boldsymbol{T}_{mn}^T \cdot Q_{mn}]^T$$

$$\boldsymbol{T}_{ij} = [T_{ij\alpha_1}, \ldots, T_{ij\alpha_4}, T_{ij\beta_1}, \ldots, T_{ij\beta_4}]^T$$

$$T_{ij\alpha_k} = \left.\frac{\partial T_{ij}}{\partial \alpha_k}\right|_{I=I_0}$$

$$T_{ij\beta_k} = \left.\frac{\partial T_{ij}}{\partial \beta_k}\right|_{I=I_0}$$

$$\boldsymbol{v} = \text{ measurement noise}$$

2. *Nonlinear Static Estimation of Ionospheric Parameters*

A nonlinear static estimation technique can be applied to the problem of fitting the ionospheric parameters to the data collected by the monitor stations.[17] We define the state $\boldsymbol{x}$ and measurement z as follows:

$$\boldsymbol{x} = \boldsymbol{I} = [\alpha_1, \ldots, \alpha_4, \beta_1, \ldots, \beta_4]^T \tag{6}$$

$$z = \boldsymbol{d} = [d_{11}, \ldots, d_{ln} \vdots, \ldots, \vdots d_{m1}, \ldots, d_{mn}]^T \tag{7}$$

The algorithm to find the solution may be formulated as follows:

1) Guess $\boldsymbol{x}$.
2) Evaluate $\boldsymbol{h}(\boldsymbol{x})$ and $\boldsymbol{H}$.
3) $\boldsymbol{P} = (\boldsymbol{M}^{-1} + \boldsymbol{H}^T\boldsymbol{V}^{-1}\boldsymbol{H})^{-1}$.
4) $\partial J/\partial \boldsymbol{x} = \boldsymbol{M}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}) - \boldsymbol{H}^T\boldsymbol{V}^{-1}[z - \boldsymbol{h}(\boldsymbol{x})] \equiv \boldsymbol{GR}$._
5) If $|\boldsymbol{GR}| \leq \epsilon$, then set $\hat{\boldsymbol{x}} = \boldsymbol{x}$ and stop. Otherwise $\boldsymbol{x} = \boldsymbol{x}$.
6) Replace $\boldsymbol{x}$ by $(\boldsymbol{x} - \boldsymbol{P} \cdot \boldsymbol{GR})$.
7) Go to (2).

B. Ephemeris and Satellite Clock Errors for Algorithms A, B, or C

The GPS navigation message broadcast by the satellites provides a means for computing the satellite positions in the WGS-84 Coordinate frame.[18] These reported positions are in error because of the limitations of the GPS control segment's ability to predict the satellite ephemeris, and potentially also because of intentional degradation of the reported parameters under SA. The GPS satellite ephemeris errors can be estimated through a network of monitor stations, by essentially using GPS upside-down. Just as a user can determine its position and clock bias based on the ranges to the known locations of four or more GPS satellites, four or more monitor stations viewing the same satellite from known locations, can be used to estimate the satellite position, clock offset, and monitor station clock offsets.

The measured pseudorange ρ_{ij}, from ith monitor station to jth GPS satellite, after being adjusted for atmospheric error and multipath error, 'is modeled by the following:

$$\begin{aligned}\rho_{ij} &= \boldsymbol{D}_{ij} \cdot \boldsymbol{e}_{ij} - B_j + b_i + n_{ij} \\ &= [(\boldsymbol{R}_j + \delta\boldsymbol{R}_j) - \boldsymbol{S}_j] \cdot \boldsymbol{e}_{ij} - B_j + b_i + n_{ij} \end{aligned} \tag{8}$$

where ρ_{ij} = measured pseudorange; $\boldsymbol{D}_{ij}$ = range vector from ith monitor station to jth satellite; $\boldsymbol{e}_{ij}$ = range unit vector from ith monitor station to jth satellite; $\boldsymbol{R}_j$ = jth satellite location calculated from the GPS message; $\delta\boldsymbol{R}_j$ = ephemeris error vector of jth satellite; $\boldsymbol{S}_i$ = known ith monitor station location; B_i = satellite clock offset; b_j = monitor station clock offset; and, n_{ij} = measurement noise. This is illustrated in Fig. 6.

Define $\boldsymbol{x}$ for all the monitor stations ($i = 1, \ldots, n$) and the GPS satellites ($j = 1, \ldots, m$) as follows:

$$\boldsymbol{x} = [\delta\boldsymbol{R}^T \quad \boldsymbol{B}^T \quad \boldsymbol{b}^T]^T \tag{9}$$

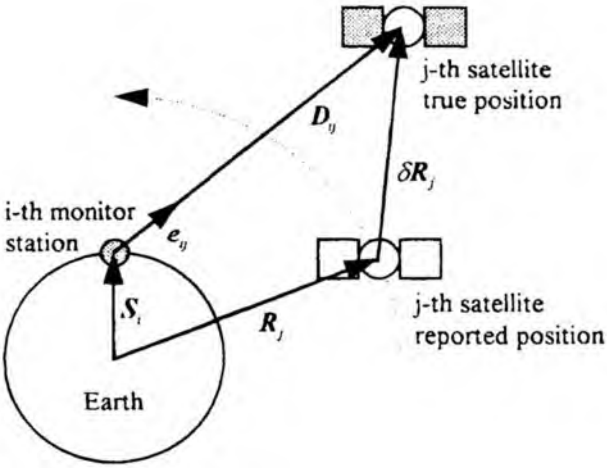


Fig. 6 GPS ephemeris errors.

where

$$\begin{aligned} \delta\boldsymbol{R} &= [\delta\boldsymbol{R}_1^T \quad \delta\boldsymbol{R}_2^T \quad \cdots \quad \delta\boldsymbol{R}_m^T]^T \\ \boldsymbol{B} &= [B_1 \quad B_2 \cdots B_{n-1}]^T \cdot \\ \boldsymbol{b} &= [b_1 \quad b_2 \cdots b_m]^T \end{aligned}$$

If we gather all the measurement Eqs. (8) for all the monitor stations ($i = 1, \ldots, n$) and the GPS satellites ($j = 1, \ldots, m$) and rearrange them, we will get a matrix equation as follows:

$$\begin{bmatrix} \boldsymbol{E}_1 & -\boldsymbol{I} & \boldsymbol{I}_1 \\ \boldsymbol{E}_2 & -\boldsymbol{I} & \boldsymbol{I}_2 \\ \vdots & \vdots & \vdots \\ \boldsymbol{E}_n & -\boldsymbol{I} & \boldsymbol{I}_n \end{bmatrix} \boldsymbol{x} = \boldsymbol{D} - \begin{bmatrix} \boldsymbol{E}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E}_2 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{E}_n \end{bmatrix} \boldsymbol{P} \tag{10}$$

where

$$\boldsymbol{E}_i = \begin{bmatrix} \boldsymbol{e}_{i1}^T & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{e}_{i2}^T & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{e}_{im}^T \end{bmatrix} \qquad (m \times 3\ m)$$

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (m \times m)$$

$$\boldsymbol{I}_i = \begin{array}{c} \begin{array}{ccccc} & & \text{(ith column)} & & \end{array} \\ \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \end{array} \qquad [m \times (n-1)] \quad (\text{for } i = 1, \ldots, n-1)$$

$$\begin{aligned} \boldsymbol{I}_n &= \boldsymbol{0} \qquad [(m \times (n-1)] \qquad (\text{for } i = n) \\ \boldsymbol{D} &= [\boldsymbol{D}_1^T \quad \boldsymbol{D}_2^T \cdots \boldsymbol{D}_n^T]^T \\ \boldsymbol{D}_i &= [\rho_{i1} \quad \rho_{i2} \cdots \rho_{im}]^T \\ \boldsymbol{P} &= [\boldsymbol{P}_1^T \quad \boldsymbol{P}_2^T \cdots \boldsymbol{P}_n^T]^T \\ \boldsymbol{P}_i &= [(\boldsymbol{R}_1 - \boldsymbol{S}_i)^T \quad (\boldsymbol{R}_2 - \boldsymbol{S}_i)^T \cdots (\boldsymbol{R}_m - \boldsymbol{S}_i)^T]^T \end{aligned}$$

In the above equations the matrix $\boldsymbol{I}_n$ is set to be $\boldsymbol{0}$ matrix because all the clock errors are relative and are estimated on the basis of the nth monitor station clock.

If we define the system matrix $\boldsymbol{H}$ and measurement z as follows

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{E}_1 & -\boldsymbol{I} & \boldsymbol{I}_1 \\ \boldsymbol{E}_2 & -\boldsymbol{I} & \boldsymbol{I}_2 \\ \vdots & \vdots & \vdots \\ \boldsymbol{E}_n & -\boldsymbol{I} & \boldsymbol{I}_n \end{bmatrix} \tag{11}$$

$$z = \boldsymbol{D} - \begin{bmatrix} \boldsymbol{E}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E}_2 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{E}_n \end{bmatrix} \boldsymbol{P} \tag{12}$$

then Eq. (10) becomes

$$z = \boldsymbol{Hx} \tag{13}$$

If the ith monitor station cannot see the jth satellite, the corresponding row element of the vector z and row vector of the matrix $\boldsymbol{H}$ in the Eq. (13) must be eliminated.

The master station uses a BLS technique to estimate the three-dimensional ephemeris error vector and clock bias for each GPS satellite within view of the network. If there are more measurements than the unknowns (three-dimensional errors, satellite clock offset, and monitor station clock offset) in the WADGPS network, the observation equation for that satellite is overdetermined, and the solution is picked to minimize the measurement residual sum of squares.

$$\boldsymbol{x} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T z \tag{14}$$

If there are fewer measurements than the unknowns (Fig. 7), the solution is underdetermined, and the optimal estimate minimizes the two-norm of the error solution.

$$\boldsymbol{x} = \boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{H}^T)^{-1} z \tag{15}$$

In the underdetermined case, the corrections for ephemeris errors and clock offsets are not accurate, but the user positioning is still accurate with these corrections because for the user, only the projection of the error correction vector on the line of sight to the satellite is important.

If the monitor stations are confined to the continental United States, users near the coastal monitor stations will be using satellites that are underdetermined, and therefore, accuracy will degrade. Consequently, we recommend locating monitor stations over a wider area than the system designed for the users.

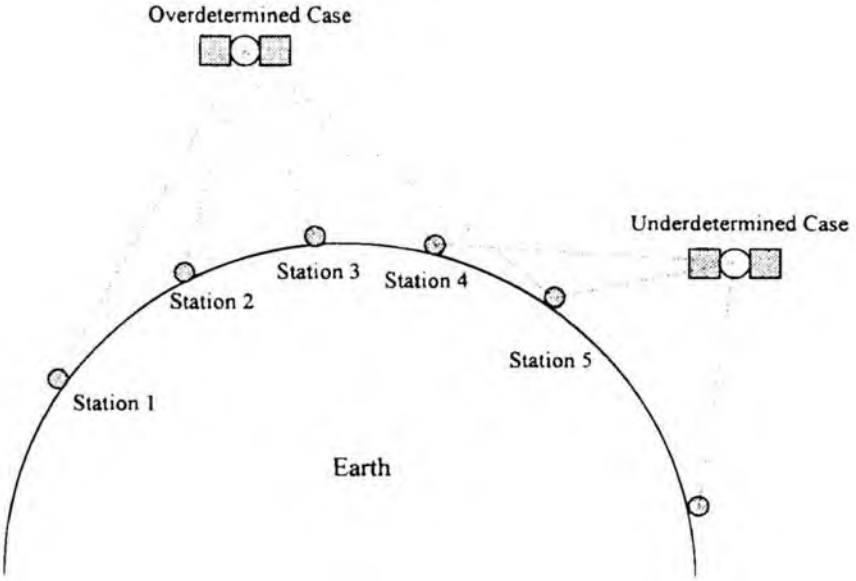


Fig. 7 Example of overdetermined and underdetermined cases of estimating ephemeris errors.

IV. Simulation of Algorithm B

The performance of the WADGPS network employing algorithm B was evaluated using a computer simulation. The simulation was run for 12 h starting at 6:00 a.m. Pacific standard time (PST).

A. Simulation Modules

The simulation is composed of four modules describing the GPS satellites, the monitor stations, the master station, and the users. A block diagram is shown in Fig. 8. The truth model error specifications are listed in Table 5.

1. GPS Satellite Module

The GPS 21 primary satellite constellation is modeled in the simulation.[19] The ephemeris reported by the GPS module to the monitor stations and the users is equal to the true ephemeris corrupted by an error vector. Each ephemeris error vector component is produced by passing white noise through a first-order shaping filter with time constant of 1800 s and standard deviation of 20 m.[20]

Each satellite clock offset is also modeled by white noise input to a first-order shaping filter, this time with $t = 200$ s,[13] and standard deviation of 30 m. These values account for possible effects caused by SA.

The ionospheric delay is modeled according to the Klobuchar model. An average ionospheric height of 350 km is assumed. In addition to the delay predicted by this model, two terms are included in our truth model to account for higher frequency

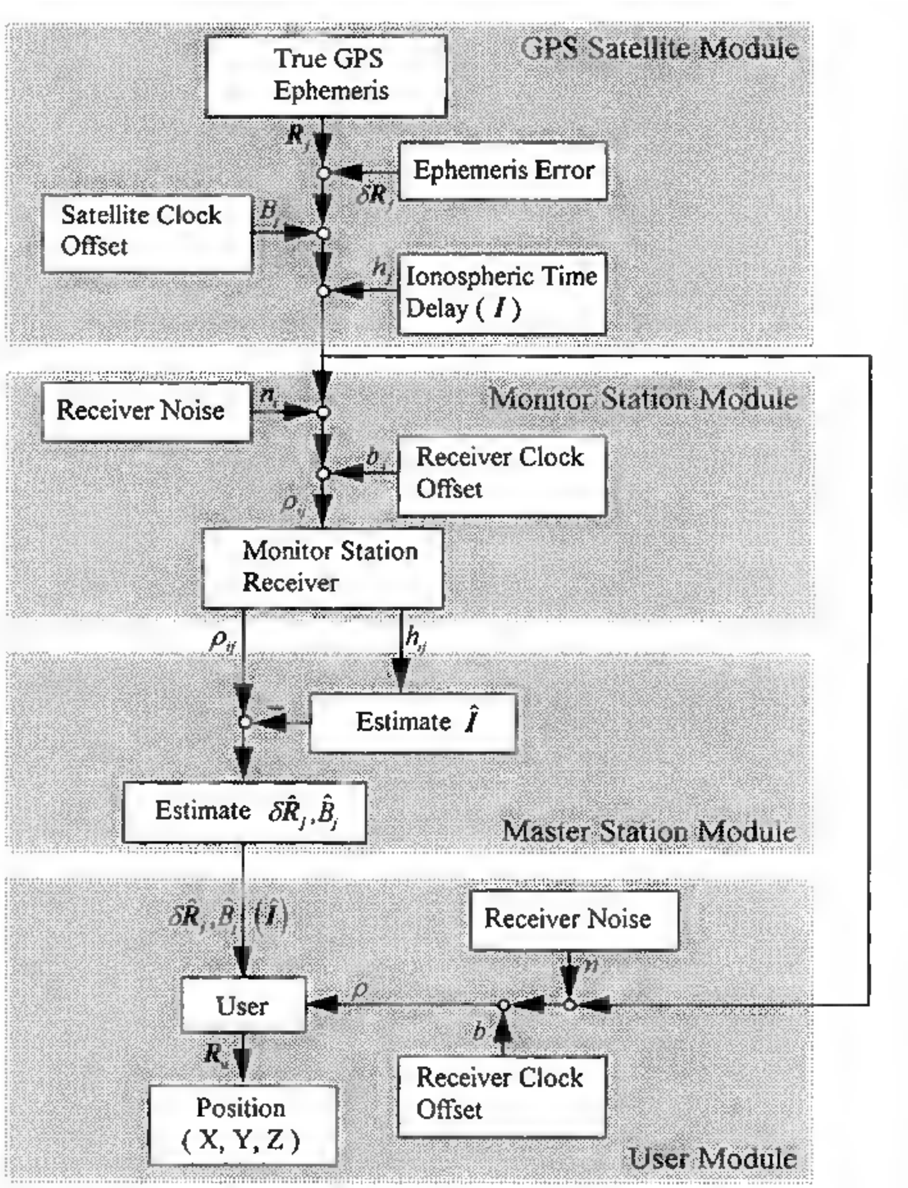


Fig. 8 Block diagram of WADGPS computer simulation.

Table 5 True model error specifications

Error source	Error model	Time constant, s	Min., m	Max., m	rms, m
Three-dimensional satellite ephemeris errors	1st-order Markov process	1800			20
Satellite clock offset	1st-order Markov process	200			30
Ionospheric time delay[a]	Klobuchar's model plus spatial sinusoidal bias and white noise	6 h	1.5 at zenith	30 at zenith	
Tropospheric error	Modeled as receiver noise[b]				
Monitor station receiver clock offset	2nd-order Markov process		$h_0 = 2.0 \times 10^{-22}$ $h_{-1} = 4.0 \times 10^{-26}$ $h_{-2} = 1.5 \times 10^{-33}$		
User receiver clock offset	2nd-order Markov process		$h_0 = 9.4 \times 10^{-20}$ $h_{-1} = 1.8 \times 10^{-19}$ $h_{-2} = 3.8 \times 10^{-21}$		
Receiver noise	White noise				0.2
Multipath	Modeled as receiver noise				

[a] Ionosphere is varying with time constant of 6 h, and 5% of one-fifth period spatial sinusoidal bias $(0.05\ \{A_1 + A_2 \cos[2\Pi(\tau - A_3)/(0.2\ A_4)]\})$ and 5% of white noise (b0.05T_{ij} × $N(0,1)$, $N(0,1)$ is Gaussian noise that has zero mean and one standard deviation) were added.

[b] A more sophisticated model is under development

ionospheric variations that have been observed in experimental data. The first is a sinusoidal error with amplitude of 5% of the cosine peak, and period of one-fifth of the cosine period of Eq. (2). The second is a random error with zero mean and standard deviation equal to 5% of the sum of the cosine terms. Ionospheric parameters I are varying from the nominal values with time constant of 6 h and result in maximum vertical ionospheric time delay 30 m and minimum, 1.5 m. A typical vertical time delay profile generated by this model is shown in Fig. 5.

2. *Monitor Station Module*

The monitor station module generates the pseudorange measurements and ionospheric delays observed by the monitor station receivers. The monitor station receiver clock offset relative to GPS time is modeled by white noise input to a second-order Markov process based on Ref. 21, as follows:

$$
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{k} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}_{k}
$$

$$
Q = E\{w \cdot w^T\} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12} & Q_{22} \end{bmatrix}
$$

$$
\begin{aligned}
Q_{11} &= \frac{h_0}{2\tau} + 2h_{-1}\tau^2 + \frac{2}{3}\pi^2 h_{-2}\tau^3 \\
Q_{12} &= 2h_{-1}\tau + \pi^2 h_{-2}\tau^2 \\
Q_{22} &= \frac{h_0}{2\tau} + 2h_{-1} + \frac{8}{3}\pi^2 h_{-2}\tau \\
h_0 &= 2.0 \times 10^{-22} \\
h_{-1} &= 4.0 \times 10^{-26} \\
h_{-2} &= 1.5 \times 10^{-33}
\end{aligned} \tag{16}
$$

where x_1 = clock offset, s; x_2 = average frequency; and, τ = sampling time. The receiver clock offset parameters h_0, h_{-1}, h_{-2} are based on a typical rubidium standard. The receiver noise is assumed to be white with zero mean and standard deviation of 0.2 m. This is based on averaging over 10 measurements at 1-s intervals.

For this simulation, 15 monitor stations are assumed, located at LORAN or VOR stations across the United States including Alaska and Hawaii. Figure 9 illustrates the location, and Table 6 lists the latitude and longitude of each station.

3. *Master Station Module*

The master station module collects inputs from the monitor station module, and implements the ionospheric and ephemeris error estimation algorithms described in Secs. III and IV. Ionospheric delay parameters, and estimated ephemeris and clock errors are provided to the user module.

4. *User Module*

The user module simulates the operation of the user receiver. The user clock error is assumed to be white noise input to a second-order Markov process based on Ref. 21, with Eq. (16) $h_0 = 9.4 \times 10^{-20}$; $h_{-1} = 1.8 \times 10^{-19}$; and $h_{-2} = 3.8 \times 10^{-21}$, where these receiver clock offset values are based on a typical quartz standard.

The receiver noise is assumed to be white with zero mean and standard deviation of 0.2 m. This is based on averaging over 10 measurements at 1-s intervals. The user applies the eight parameters in the Klobuchar model to the raw pseudorange and adjusts the ephemeris parameters received from the GPS satellite module by the correction vector sent by the master station. Then the user forms a least-squares position solution using measurements to all the satellites within his field of view. The performance of the WADGPS is evaluated by

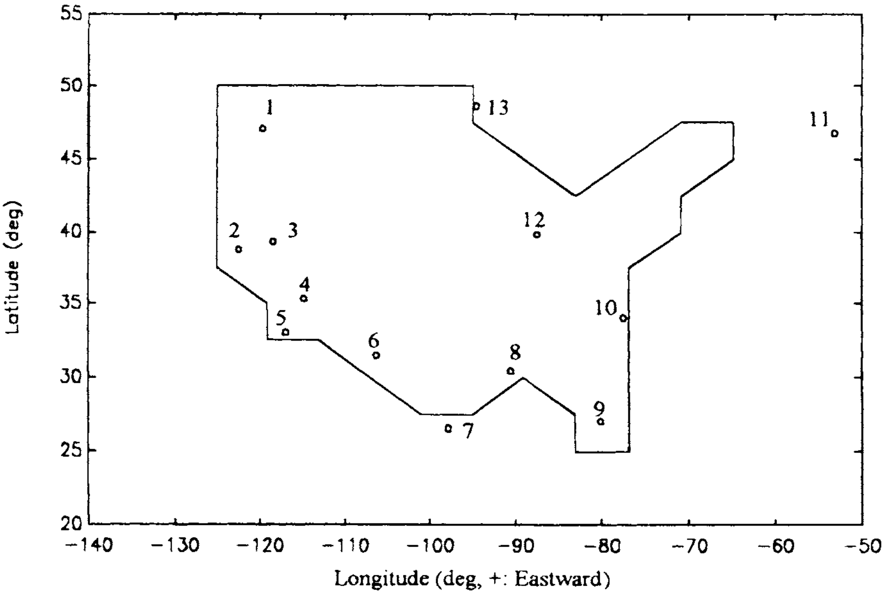


Fig. 9 Locations of monitor stations in United States (Narrow Cape and Upolo Point are not shown on the map).

Table 6 Locations of monitor stations

Location	Latitude	Longitude	LORAN site	VOR site
1) George, WA	47:04 N	119:45 W	√	
2) Middletown, CA	38:47 N	122:30 W	√	
3) Fallon, NV	39:33 N	115:50 W	√	
4) Searchlight, NV	35:19 N	114:48 W	√	
5) San Diego, CA	33:00 N	117:00 W		√
6) El Paso, TX	31:30 N	106:20 W		√
7) Raymondville, TX	26:32 N	97:50 W	√	
8) Grangeville, LA	30:43 N	90:50 W	√	
9) Jupiter, FL	27:02 N	80:07 W	√	
10) Carolina Beach, NC	34:04 N	77:55 W	√	
11) Cape Race, Newfoundland, Canada	46:47 N	53:10 W	√	
12) Dana, IN	39:51 N	87:29 W	√	
13) Baudette, MN	48:37 N	94:33 W	√	
14) Narrow Cape, Kodiak Is., AK (not shown on the map)	57:26 N	152:22 W	√	
15) Upolo Pt., HI (not shown on the map)	20:15 N	155:23 W	√	

comparing the error in this solution to the error that would have been obtained by using the raw measurements directly.

A typical user who would benefit from the WADGPS, has a single-frequency, C/A-code GPS receiver and a quartz oscillator. Eighty-one stationary users are modeled at locations distributed uniformly across the United States (Fig. 10). All users are assumed to have an elevation mask angle of 6.5 deg, which is typical antenna visibility for an aircraft.

B. Ionospheric Error Estimation Results

The first step in evaluating the performance of the WADGPS is to see how well it determines the ionospheric errors. Figures 11 and 12 show contour plots of the actual and estimated vertical ionospheric delays superimposed on a map of the United States. These represent snapshots from the 12-h simulation at 5 p.m. PST and 2 p.m. PST, respectively. The contour lines are labeled by the ionospheric delay in 3-m increments. Note that the actual values of the ionospheric delay increase from east to west as we get closer to local noon.

In the 5 p.m. plot, we can see that the nonlinear static estimator algorithm does well at estimating the delay because the estimated contours are within 1.5 m of the true error contours. The performance results are summarized in Table 7.

Figure 11 also shows the improvement in ionospheric delay estimates toward the center of the country. This can be attributed to the larger number of monitor stations that can observe satellite signals passing through the central region as compared to satellites in the far eastern or western parts of the sky.

Figure 12 shows similar results for 2 p.m. PST. In this case, however, we notice local contours of varying heights that are not estimated. These small areas of variation in the ionospheric delay are generated by the random noise introduced in the truth model. Because the standard deviation of this variation is set at 5% of the nominal value for the local time of day, the maximum random error is as

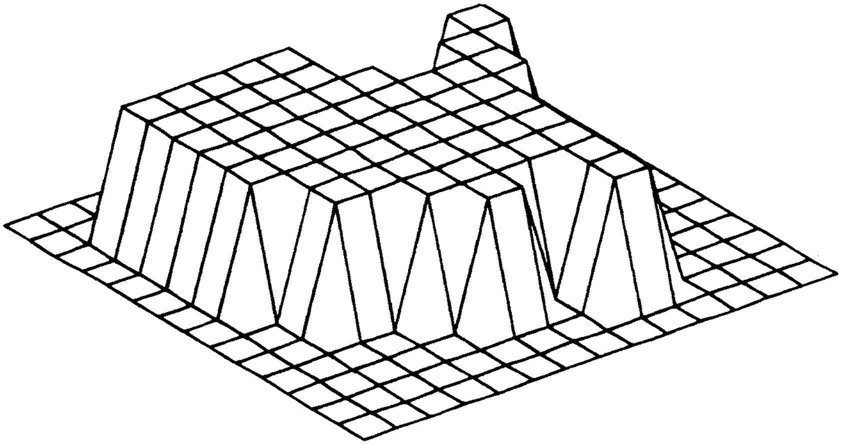


Fig. 10 Mesh plot of the continental United States. (Each point on the grid represents one of the 81 simulated user's positions.)

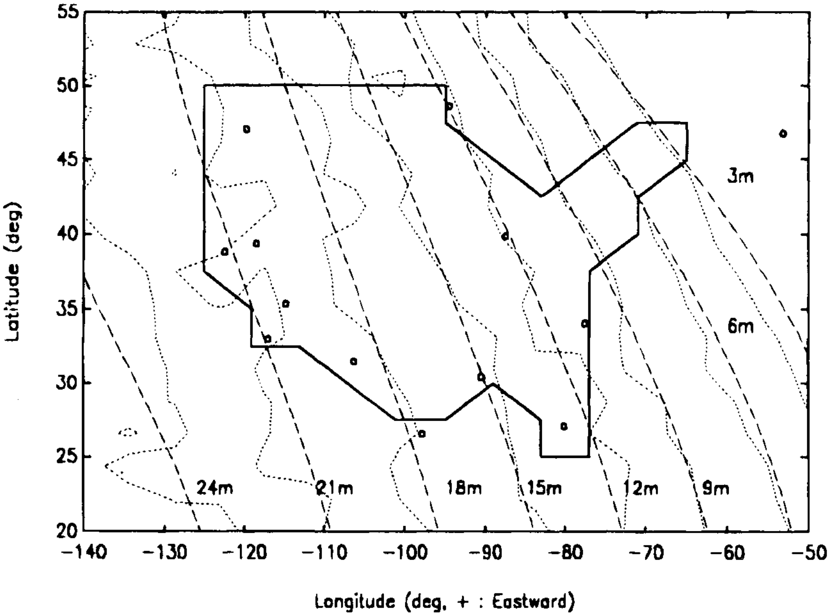


Fig. 11 Ionospheric time delay estimates (5:00 p.m. PST). (This figure shows the vertical ionospheric delay in meters as generated by the truth model and the NSE. A map of the United States and the monitor station locations is also shown. The dotted line is the true delay contour; the dashed line is the NSE estimate.)

much as 4.5 m at 2 p.m. There is no way for our rather sparse WADGPS network model to estimate these simulated random, high-frequency, localized components of the ionospheric error.

C. Navigation Performance

The objective of the WADGPS system is to improve navigation accuracy for users. The simulation results indicate that this goal can be achieved using the proposed system.

Figures 13–18 provide a very compact summary of the simulation results. In these mesh plots, each grid point within the outline of the United States represents 1 of the 81 user locations we considered. The height of the grid point above the surface corresponds to the magnitude of the error at the grid location. These heights reflect rms or maximum error for that user over the entire 12-h simulation period.

Figures 13 and 16 show the uncorrected navigation performance of typical across the United States. The rms of the positioning errors for all user locations, over the entire 12-h period, is 82 m in vertical direction, and 42 m in the horizontal direction. As we might expect, the error magnitudes are fairly uniform over the entire area. As is common in GPS navigation, the vertical error is approximately twice as large as the horizontal error because of the larger geometrical dilution of precision in the vertical direction (VDOP).

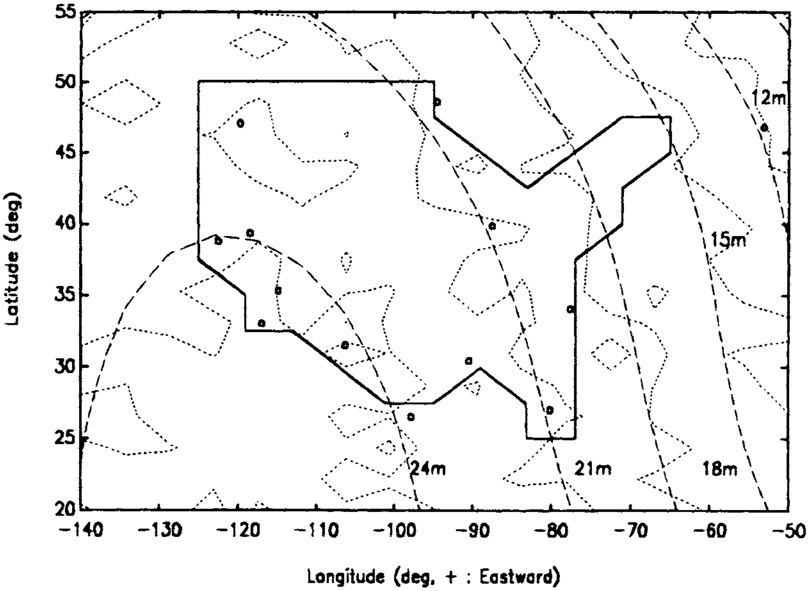
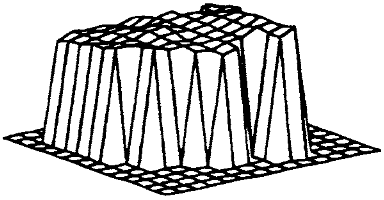


Fig. 12 Ionospheric time delay estimates (2:00 p.m. PST). (This figure shows the vertical ionospheric delay in meters as generated by the truth model and the NSE. A map of the United States and the monitor station locations is also shown. The dotted line is the true delay contour; the dashed line is the NSE estimate.)

Table 7 Root-mean-square and maximum errors in ionospheric estimates

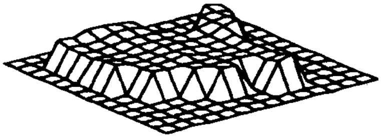
PST local time		7:00–8:00 a.m.	10:00–11:00 a.m.	1:00–2:00 p.m.	4:00–5:00 p.m.
Nonlinear static	max,[a] m	1.0	1.4	1.3	1.3
Estimation	rms,[b] m	0.3	0.5	0.5	0.4

[a] max[abs(z_i − $\hat{z}_i$)] where z_i is the true vertical ionosphere measurement and $\hat{z}_i$ the estimated value.
[b] rms(z_i)



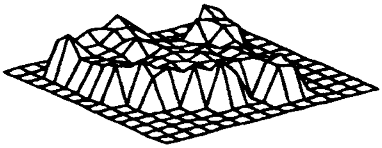
max of rms=93.1m
continent rms=82.1m

Fig. 13 Root-mean-square value of stand-alone GPS vertical positioning errors.



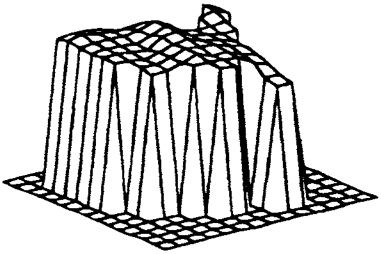
max of rms=1.9m
continent rms=1.5m

Fig. 14 Root-mean-square value of WADGPS vertical positioning errors.



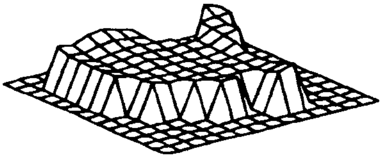
max of max=9.1m
continent rms of max=6.3m

Fig. 15 Maximum WADGPS vertical positioning errors.



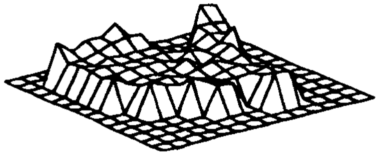
max of rms=45.4m
continent rms=42.3m

Fig. 16 Root-mean-square value of stand-alone GPS horizontal positioning errors.



max of rms=1.5m
continent rms=1.1m

Fig. 17 Root-mean-square value of WADGPS horizontal positioning errors.



max of max=6.0m
continent rms of max=3.6m

Fig. 18 Maximum WADGPS horizontal positioning errors.

Figures 14–15 and 17–18 show the significantly improved navigation accuracy achieved using WADGPS with the NSE ionospheric estimation algorithm and ephemeris and clock bias BLS algorithm. Continental rms averages of the vertical and horizontal position errors have been reduced from 45.4 m and 42.3 m to 1.5 m and 1.1 m, respectively. Figures 15 and 18 show the maximum values of the errors. The largest vertical and horizontal errors anywhere in the United States over the entire 12-h period are 9.1 m and 6.0 m, respectively.

One of the most striking features of the plots of WADGPS corrected errors is the concave shape of the error mesh. In general, the navigation performance in the center of the United States is better than along the coasts. This is because the satellites viewed by users in this region are also visible from a larger number of monitor stations, and with better geometry [lower geometric dilution of precision (GDOP)] than their coastal counterparts.

One exception to this observation occurs in the southwestern United States. This region exhibits better positioning accuracies than other edges of the country because of the high density of monitor stations (Figs. 14–15 and 17–18). Likewise, the north central part of the country is noticeably worse than average because of the relative sparsity of monitor stations.

D. Summary of Results

A 12-h simulation was run starting at 6:00 a.m. PST and ending at 6:00 p.m. PST. The NSE technique was used to determine the ionospheric delay parameters. Table 8 summarizes the simulation results.

Simulation results over a 12-h period indicate that stand-alone GPS positioning errors can be reduced by over 95% using WADGPS without degradation caused by separation between the monitor station and the users. These results indicate that WADGPS can provide accurate ionospheric delay estimates as well as positioning errors comparable to local area differential GPS operations.

V. Test Using Field Data to Evaluate Algorithm C

Previously collected field test data (from Dec. 10, 1992 to Feb. 12, 1993 using the GPS Global Tracking Network) were processed to evaluate WADGPS performance. The GPS Global Tracking Network has more than 30 sites distributed worldwide, which are equipped with P-code receivers and of which locations are known to within a few centimeters, and the Jet Propulsion Laboratory (JPL)

Table 8 Summary of positioning errors

	Nonlinear static estimation			
	Vertical		Horizontal	
	Maximum[a]	rms[b]	Maximum[a]	rms[b]
Root-mean-square of stand-alone positioning error, m	93.1	82.1	45.4	42.3
Root-mean-square value of WADGPS positioning . error, m	1.9	1.5	1.5	1.1
WADGPS error/stand-alone error, %	2.6	1.9	3.0	2.1
Max. value of WADGPS positionng error, m	9.1	6.3	6.0	3.6

[a] Maximum for the continental United States.
[b] Root-mean-square for the continental United States.

collects data from all the network sites. Sampling time for the most sites was 30 s, and the available measurements were L_1 and L_2 P-code pseudoranges and L_1 and L_2 continuous carrier phases. C/A-code pseudoranges and Doppler measurements were not available and SA was on during the field test. Because meteorological data were not available, the temperature, pressure, and humidity of each site were inferred by location and time of day.

Algorithm C was used for this test. It provides the best accuracy because the user uses a dual-frequency receiver to measure ionospheric time delays very accurately. As such, the results presented here provide a estimate of the best possible performance of WADGPS.

A. Locations of the Receiver Sites

Among over 30 Rogue receiver sites, 7 were picked in North America and Hawaii for the field test. Six sites were chosen as monitor stations for WADGPS because their sites are evenly distributed and their geometry constitutes a rough circle. One site, ALBH (Albert Head, Canada), was picked as user because it is near the center of the circle, and therefore, could demonstrate the potential of WADGPS. The minimum baseline between the user (ALBH) and a monitor station (JPLM) was 1632 km. The locations of the receiver sites are listed in Table 9 and the corresponding map is in Fig. 19.

All the coordinates of the receiver locations were given in the International Terrestrial Reference Frame (ITRF) 91 coordinate frame instead of the WGS-84 frame in which the GPS ephemeris is computed. The ephemeris errors in the ITRF91 frame are different from those calculated in the WGS-84 frame, but using the same coordinate frame for the monitor stations and the user avoids unexpected positioning errors.

B. Test Results

As soon as a set of data from monitor stations became available, the master station estimated the ephemeris errors and satellite clock errors and transmitted

Table 9 Locations of the P-code Receiver Sites

Site type	Station ID	City	Nation	Latitude deg	Longtitude deg	Baseline from, ALBH, km
Monitor	ALGO	Algonquin	Canada	46.0 N	78.0 W	3363
Monitor	FAIR	Fairbanks, AK	USA	65.0 N	147.5 W	2318
Monitor	JPLM	Pasadena, CA	USA	34.1	118.1 W	1632
Monitor	KOKB	Kokee, HI	USA	22.1 N	159.7 W	4245
Monitor	RCM2	Richmond, FL	USA	25.6 N	80.4 W	4414
Monitor	YELL	Yellowknife	Canada	62.5 N	114.5 W	1661
User	ALBH	Albert Head	Canada	48.4 N	123.5 W	0

the error corrections to users. All the data were postprocessed as if it were in real time. Normally it took about 3–4 s to compute the error corrections for each epoch using PC-486/25 computer. However, the data were applied without delay.

A total of six satellites were in view from ALBH during the field test. Typically six to nine satellites will be seen from a receiver when GPS is in full operation 1994. Figure 20 shows azimuth vs. elevation plot during the test period.

Only six monitor stations were used in this test, which is considerably fewer than 15 stations used in the simulations (Sec. IV). Not all the satellites were in view from more than 4 monitor stations, which is the underdetermined case. In that case, the estimates of the ephemeris errors and clock offsets were not accurate, but the user positioning was still accurate with the WADGPS corrections.

We tested a total of 12 days, and the results are summarized in Table 10. We show the worst case (1/23/93) in Figs. 21–23, and one of the best cases (1/13/93) in Figs. 24–26.

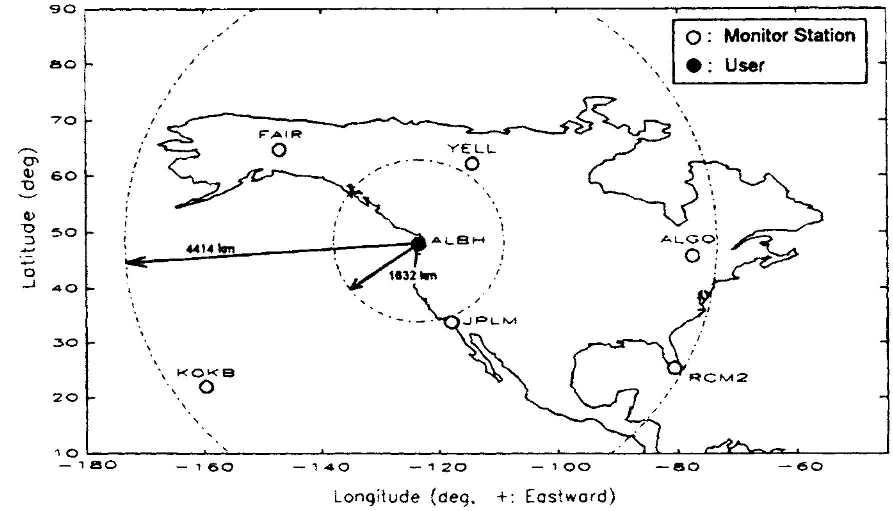


Fig. 19 Location map of the P-code receiver sites.

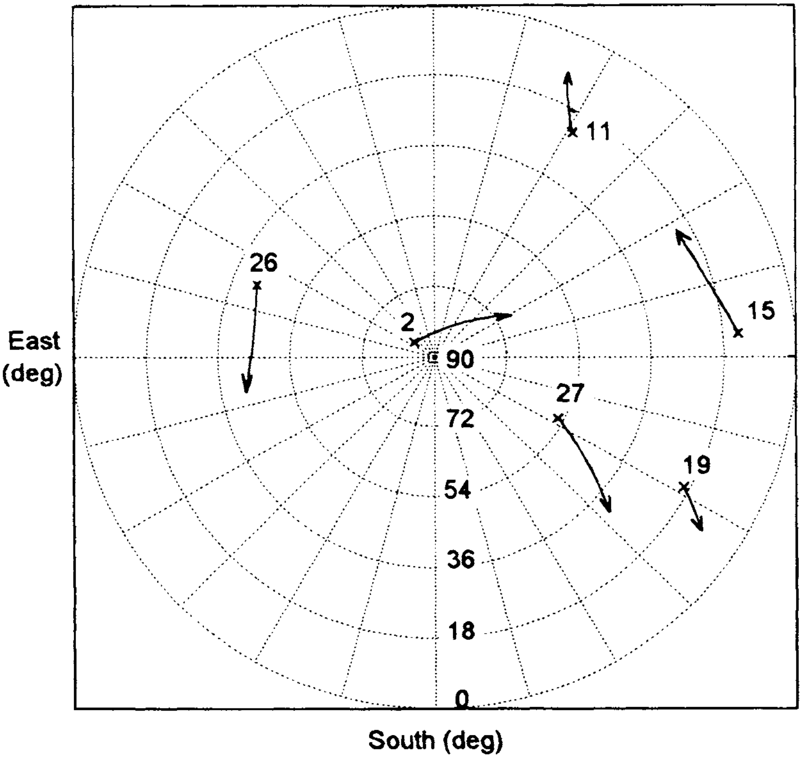


Fig. 20 Azimuth vs elevation plot (ALBH, 1/23/93).

Table 10 Summary of navigation errors at ALBH using Algorithm C and zero latency (1632 km baseline, 12/10/92–2/12/93)

Date	GPS time	GPS, three-dimensional rms, m	WADGPS, three-dimensional	WADGPS/ GPS, %
12/10/92	4:10 a.m.–4:36 a.m.	57.0	0.71	1.2
1/6/93	4:34 a.m.–5:00 a.m.	55.4	0.84	1.5
1/11/93	3:12 a.m.–3:42 a.m.	52.9	1.11	2.1
1/12/93	3:15 a.m.–3:38 a.m.	50.1	1.32	2.6
1/13/93	3:38 a.m.–4:10 a.m.	62.0	0.86	1.4
1/14/93	3:34 a.m.–4:05 a.m.	48.5	1.51	3.1
1/23/93	3:00 a.m.–3:55 a.m.	65.7	2.18	3.3
1/29/93	3:01 a.m.–3:47 a.m.	81.3	1.45	1.8
1/30/93	3:07 a.m.–3:55 a.m.	51.3	1.99	3.9
1/31/93	3:03 a.m.–3:47 a.m.	68.5	1.58	2.3
2/11/93	3:22 a.m.–3:59 a.m.	73.9	0.94	1.3
2/12/93	3:20 a.m.–3:55 a.m.	67.5	0.94	1.4
Total average		61.2	1.29	2.2

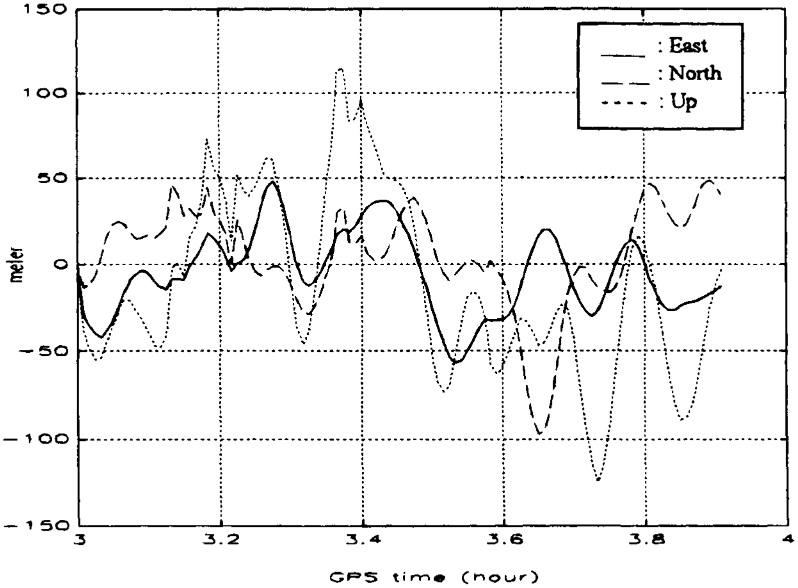


Fig. 21 Stand-alone user positioning error (ALBH, 1/23/93).

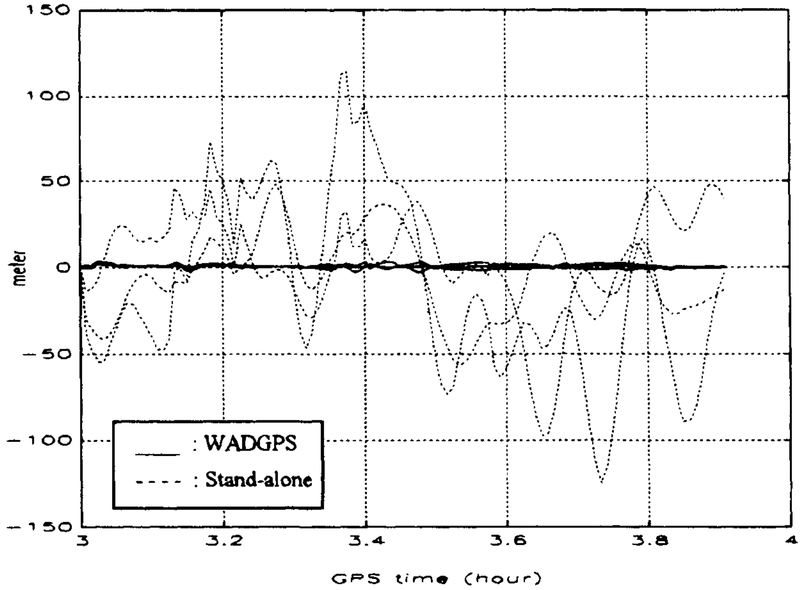


Fig. 22 Stand-alone vs WADGPS user positioning error (ALBH, 1/23/93).

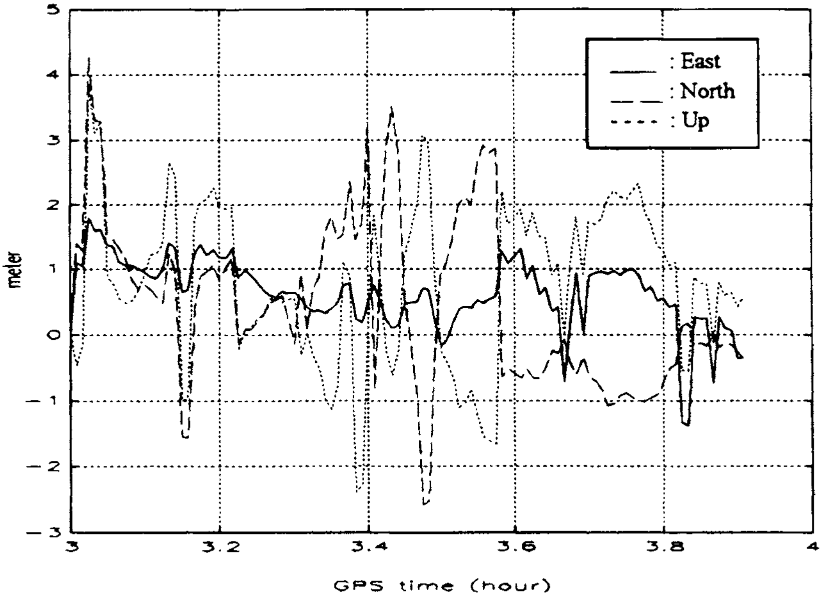


Fig. 23 WADGPS user positioning error (ALBH, 1/23/93).

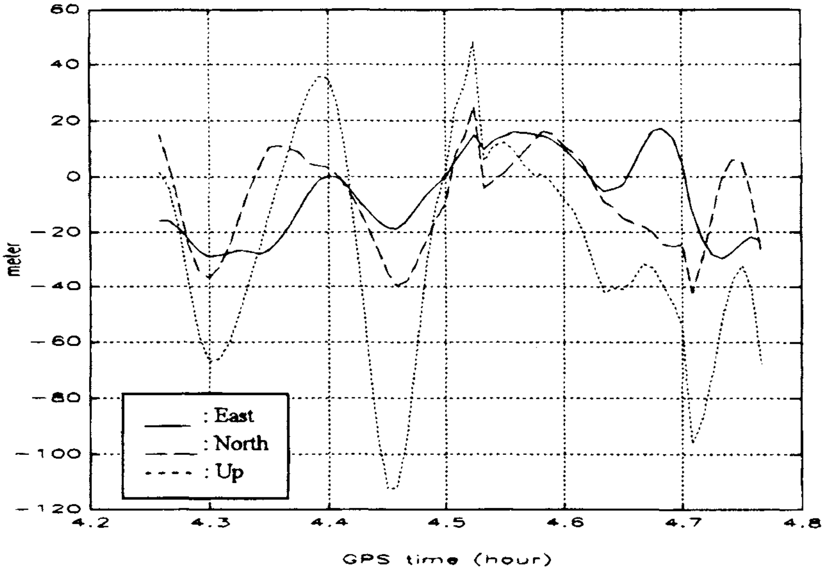


Fig. 24 Stand-alone user positioning error (ALBH, 1/13/93).

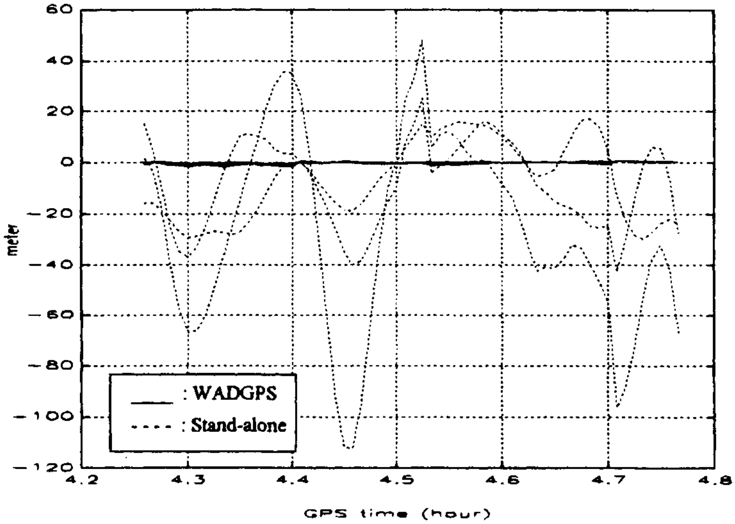


Fig. 25 Stand-alone vs WADGPS user positioning error (ALBH, 1/13/93).

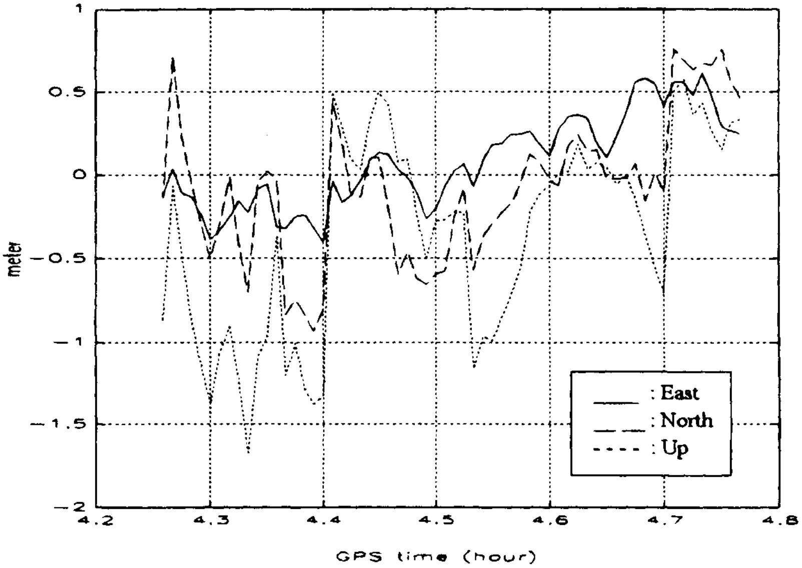


Fig. 26 WADGPS user positioning error (ALBH, 1/13/93).

Figures 21 and 24 show the stand-alone user positioning errors at night at ALBH on January 23, 1993 and January 13, 1993, respectively. Their rms three-dimensional positioning errors for stand-alone user are 65.7 m and 62 m, which is typical under SA. The navigation errors contain large oscillations with time constants of approximately 2 min, indicating the presence of SA.

Figures 22 and 25 show the significantly improved WADGPS vs. stand-alone GPS navigation errors, and Figs. 23 and 26 show the only WADGPS positioning errors with much smaller scale. The three-dimensional rms positioning errors of WADGPS user for the days, January 23 and January 13, are 2.18 m and 0.86 m, respectively. The shortest baseline from the monitor station to ALBH is 1632 km, which is very long, but WADGPS provides nonspatially degrading error corrections to user.

Overall, a normal GPS three-dimensional positioning error, 61.2 m, is reduced to 1.29 m using WADGPS. In other words about 98% of the normal GPS three-dimensional positioning error can be eliminated using WADGPS.

Two important error sources for WADGPS positioning errors are multipath and tropospheric refraction. Multipath can be reduced by using a choke ring antenna and the Hatch/Eshenbach filter but cannot be totally eliminated. Better tropospheric models can reduce the tropospheric error, but a residual error still exists.

C. Latency and Age Concern

The results of the test showed that WADGPS can achieve navigation accuracy on the order of one meter even for a 1632-km baseline with zero latency, which is not achievable for local area DGPS. However, in practice, it is impossible for users to apply the error corrections at the same epoch at which they are estimated in the master station. We define latency as the time taken to estimate the correction parameters plus the time spent for the WADGPS error corrections to arrive at the users via geosynchronous satellite. Actually, users have to use the old correction message until the new correction message arrives. So we define age, total time delay, as latency plus the time interval from when the old correction message arrived till when users apply a new correction message. Figure 27 shows the definition of latency and age.

Usually the maximum age for 5–10 s of latency is 10–20 s. The major source of the error caused by the time delay is SA, which has a 2–3 min time constant and is the fastest changing error source.

The GPS Global Tracking Network data taken on December 10, 1992 was used to investigate the latency effects on WADGPS navigation accuracy. Because

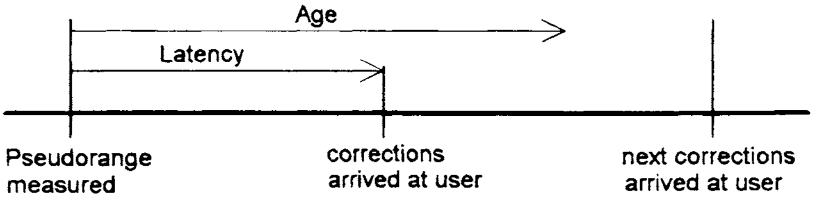


Fig. 27 Definition of latency and age.

the sampling time of the data was 30 s, the tests were repeated with latencies of 30–180 s with 30-s increments. We used two different error correction techniques. The first technique keeps the error corrections constant and after a certain time delay applies these fixed corrections to the users. The second technique estimates the error–rate corrections as well as error corrections and uses both terms to predict an error correction for the user.

Fig. 28 gives the effect of the different latency values on the three-dimensional positioning accuracies, and Fig. 29 shows the effect of the latency on an enlarged scale. Table 11 summarizes the effect of latency on WADGPS navigation accuracy.

Figure 29 indicates that a 1.5–3 m three-dimensional rms positioning accuracy can be achieved with 5–10 s of latency, which corresponds to 10–20 s of age, if WADGPS is used with error–rate corrections (rather than with constant error corrections). On the worst case day (1/23/93), 3–5 m three-dimensional rms positioning accuracy can be achieved with 5–10 s of latency using WADGPS.

VI. Conclusion

This chapter has introduced the WADGPS concept, which includes one or more master stations, monitor stations, and a broadcast system. Each functional component of WADGPS was described. Then the user message content and its format were proposed, and a listing of the error budget for WADGPS was given. The procedures used at the master station for estimating ionospheric time delay parameters and ephemeris and satellite clock errors were discussed in detail.

Three different real-time WADGPS algorithms have been proposed, which depend on receiver type in the monitor stations and users and broadcasting

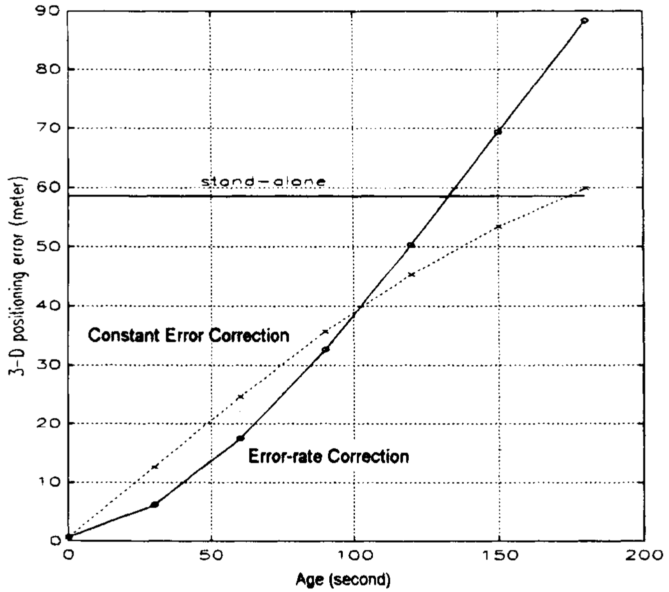


Fig. 28 Age effect on WADGPS (ALBH, 12/10/92).

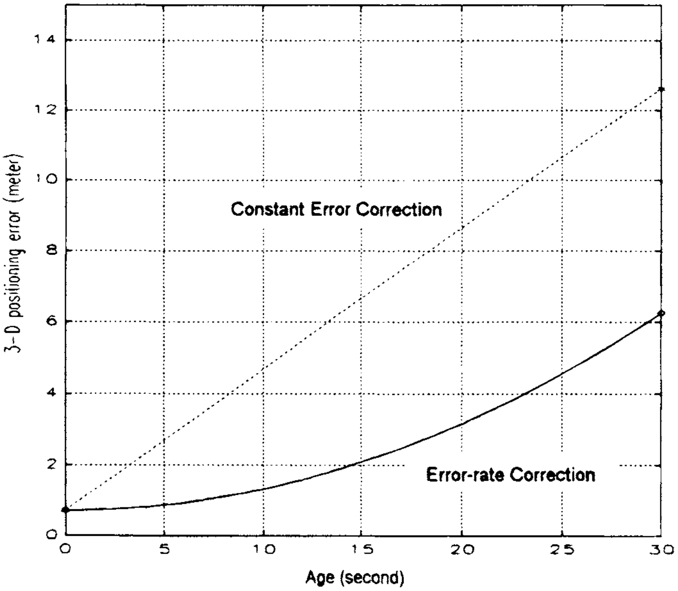


Fig. 29 Age effect on WADGPS (ALBH, 12/10/92).

parameters from the master station to the users. Algorithm B was chosen to simulate WADGPS because it includes the effect of the ionospheric time delay in WADGPS navigation accuracy. algorithm C, which estimates only three-dimensional ephemeris errors and satellite clock offset including SA assuming that users can measure the ionospheric time delays by themselves, was evaluated using field data because it provides a best possible bound for WADGPS performance.

The GPS Global Tracking Network data collected by P-code receivers were used to evaluate algorithm C. Six monitor stations that provide good geometry and one user site near the center of the network were picked in North America and Hawaii to demonstrate WADGPS performance. A batch least squares and minimum norm solution were used in the estimation.

The test results indicate that stand-alone GPS positioning error, which is 61.2 m, can be reduced to on the order of a meter (1632-km baseline) using algorithm C with zero latency. However, latency will degrade the WADGPS navigation

Table 11 Three-dimensional WADGPS rms positioning errors in the existence of latency (age) for two different error correction techniques (ALBH, 12/10/92)

Latency, s	0	30	60	90	120	150	180
Constant error correction	0.71	12.63	24.64	35.68	45.38	53.64	59.85
Error-rate correction	0.71	6.25	17.49	32.67	50.33	69.33	88.33

accuracy, but a study of the latency showed that 2–4 m of three-dimensional rms positioning accuracy can be achieved with 5–10 s of latency, which corresponds to 10–20 s of age, if WADGPS is used with error–rate corrections. Of course, algorithm C assumes that users have dual-frequency receivers or can estimate the ionospheric time delays accurately. If this assumption is not true, the positioning accuracy will be further degraded, but still should be better than 10 m.

The prediction of WADGPS above has been verified in flight trials by Stanford Telecommunication and the FAA[22,33] so that WADGPS will be used as the basis of the Wide Area Augmentation System (WAAS) to be operational by 1997, which is detailed in the next chapter.

References

[1] Chou, H., "An Adaptive Correction Technique for Differential Global Positioning System," Ph.D. Dissertation, Stanford University, Stanford, CA, June, 1991.

[2] Kremer, G. T., Kalafus, R. M., W. Loomis, P. V., and Reynolds, J. C., "The Effect of Selective Availability on Differential GPS Corrections," *Navigation,* Vol. 37, No. 1, 1990, pp. 39–52.

[3] Kee, C., Parkinson, B. W., and Axelrad, P., "Wide Area Differential GPS," *Navigation,* Vol. 38, No. 2, 1991.

[4] Kee, C., and Parkinson, B. W., "Algorithms and Implementation of Wide Area Differential GPS," *Proceedings of ION GPS-92* (Albuquerque, NM), ION, Washington, DC, Sept. 16–18, 1992, pp. 565–572.

[5] Kee, C., and Parkinson, B. W., "High Accuracy GPS Positioning in the Continent: Wide Area Differential GPS," Differential Satellite Navigation Systems 93 (DSNS-93) Conference, Amsterdam, The Netherlands, April 1993.

[6] Kee, C., and Parkinson, B. W., "Static Test Results of Wide Area Differential GPS," *Proceedings of ION GPS-93* (Salt Lake City, UT), ION, Washington, DC, Sept. 22–24, 1993.

[7] Brown, A., "Extended Differential GPS," *Navigation,* Vol. 36, No. 3, 1989.

[8] Loomis, P. V. W., Denaro, R. P., and Saunders, P., "Worldwide Differential GPS for Space Shuttle Landing Operations," *IEEE PLANS '90, IEEE Position, Location, and Navigation Symposium* (Las Vegas, NV, IEEE, New York, March, 1990.

[9] Cohen, C., Pervan, B., and Parkinson, B. W., "Estimation of Absolute Ionospheric Delay Exclusively though Single Frequency GPS Measurements," *Proceedings of ION GPS-92* (Albuquerque, NM), ION, Washington, DC, Sept. 16–18, 1992, pp. 325–330.

[10] Xia, R. "Determination of Absolute Ionospheric Error using a Single Frequency GPS Receiver," *Proceedings of ION GPS-92,* (Albuquerque, NM), ION, Washington, DC, Sept. 16–18, 1992, pp. 483–490.

[11] Weber, L., and Tiwari, A., "Performance of a FM Sub-Carrier (RDS) Based DGPS System," *Proceedings of ION GPS-93,* (Salt Lake, UT), ION, Washington, DC, Sept. 22–24, 1993, pp. 1285–1292.

[12] Kalafus, R. M., "Special Committee 104 Recommendations for Differential GPS Service," *Navigation,* Vol. 3, 1986, pp. 101–116.

[13] Chou, H., "An Anti-SA Filter for Non-differential GPS Users," *Proceedings of ION GPS-90* (Colorado Springs, CO), ION, Washington, DC, Sept. 19–21, 1990.

[14] El-Arini, M., O'Donnell, P. A., Kellam, P. M., Klobuchar, J. A., Wisser, T. C., and Doherty, P. H., "The FAA Wide Area Differential GPS (WADGPS) Static Ionospheric

Experiment," *National Technical Meeting of the Institute of Navigation,* (San Francisco, CA), ION, Washington, DC, Jan. 1993, pp. 485–496.

[15]Klobuchar, J., Air Force Geophysical Laboratory, "Design and Characteristics of the GPS Ionospheric Time Delay Algorithm for Single Frequency Users," *IEEE PLANS '86, Position, Location, and Navigation Symposium* (Las Vegas, NV), IEEE, New York, Nov. 4–7, 1986.

[16]Stephens, S. G., and Feess, W. A., "An Evaluation of the GPS Single Frequency User Ionospheric Time Delay Model," *IEEE PLANS '86, Position, Location, and Navigation Symposium,* (Las Vegas, NV), IEEE, New York, Nov. 4–7, 1986.

[17]Bryson, A. E., "Optimal Estimation and Control Logic in the Presence of Noise" (Lecture Notes, Stanford Course AA278B), Spring 1989.

[18]"Navstar GPS Space Segment/Navigation User Interfaces," Rockwell International Corporation, ICD-GPS-200, Nov. 30, 1987.

[19]Green, G. B., Massatt, P. D., and Rhodus, N. W., "The GPS 21 Primary Satellite Constellation," *Navigation,* Vol. 36, No. 1, Spring, 1989.

[20]Wells, D., "Guide to GPS Positioning," Canada GPS Associates, 1987.

[21]Van Dierendonck, A. J., McGraw, J. B., and Brown, R. G., "Relationship between Allan Variances and Kalman Filter Parameters," *Proceedings of the 16th annual PTTI Applications & Planning Meeting,* Maryland, Nov. 1984.

[22]Lage, M. E., and Elrod, B. D., "The FAA's WIB/WDGPS Testbed and Recent Test Results," *Proceedings of ION GPS-93,* (Salt Lake, UT), ION, Washington, DC, Sept. 16–18, 1993, pp. 487–493.

[23]Loh, R., Persello, F., and Wollschleger, V., "FAA's Wide Area Augmentation System (WAAS) Summary of Ground & Flight Test Data," *Proceedings of ION GPS-94* (Salt Lake, UT), ION, Washington, DC, Sept. 16–18, 1994, pp. 99–106.

Chapter 4

Wide Area Augmentation System

Per K. Enge*
Stanford University, Stanford, California 94305
and
A. J. Van Dierendonck†
AJ Systems, Los Altos, California 94024

I. Introduction

IN time, the Global Positioning System (GPS) will be used for a wide variety of aircraft operations. However, aircraft use of any satellite-based navigation system raises significant concern with respect to integrity (hazardous but undetected faults), reliability (continuity of service), time availability, and accuracy. After all, a single satellite malfunction would affect users over a huge geographic area. A navigation system with integrity warns its users if position errors may be greater than a prespecified "alarm limit." Clearly, radionavigation systems used by aviators must have integrity, and the integrity requirement depends on whether the system is the primary navigation aid or supplements another system.

The wide area augmentation system (WAAS) is a safety-critical system consisting of a signal-in-space and a ground network to support enroute through precision approach air navigation. It is designed to augment GPS so that it can be used as the primary navigation sensor. The WAAS augments GPS with the following three services: a ranging function, which improves availability and reliability; differential GPS corrections, which improve accuracy; and integrity monitoring, which improves safety.

The WAAS is shown in Fig. 1. As shown, the WAAS broadcasts GPS integrity and correction data to GPS users and also provides a ranging signal that augments GPS. In the near future, the WAAS signal will be broadcast to users from geostationary satellites. Inmarsat-3 satellites will be the first to carry the WAAS navigation payload, and they are scheduled to be launched in 1996.

*Professor (Research), Department of Aeronautics and Astronautics.

†Systems Consultant.

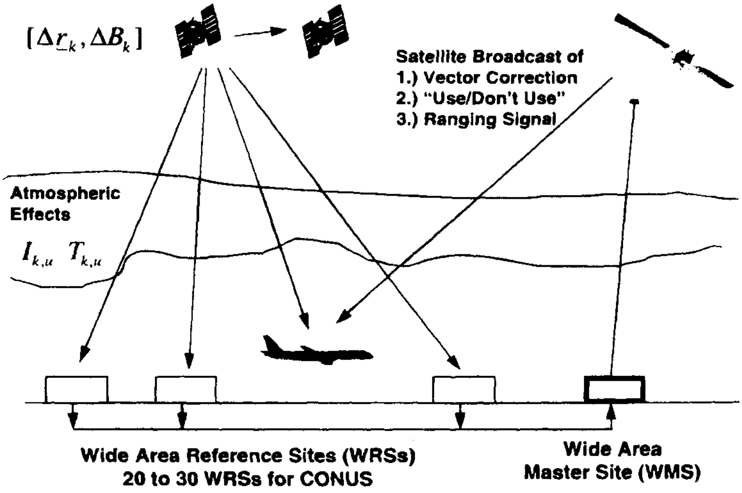


Fig. 1 Wide area augmentation system.

The WAAS ranging signal is GPS-like and will be received by slightly modified GPS receivers. More specifically, it will be at the GPS L_1 frequency and will be modulated with a spread spectrum code from the same family as the GPS C/A-codes. The code phase and carrier frequency of the signal will be controlled so that the WAAS satellites will provide additional range measurements to the GPS user. The WAAS signal will also carry data that contain differential corrections and integrity information for all GPS satellites as well as the geostationary satellite(s).

The ground network shown in Fig. 1 develops the differential corrections and integrity data broadcast to the users. Wide area reference stations (WRS) are widely dispersed data collection sites that receive and process signals received from the GPS and geostationary satellites. An example network of WRS for the United States is shown in Fig. 2. The WRS forward their data to data-processing sites referred to as wide area master stations (WMS or central processing facilities).

The WMS process the raw data to determine integrity, differential corrections, residual errors, and ionospheric delay information for each monitored satellite. They also develop ephemeris and clock information for the geostationary satellites. All these data are packed into the WAAS message, which is sent to navigation Earth stations (NES). The NES uplink this message to the geostationary satellites, which broadcast the "GPS-like" signal described earlier.

Taken together, the differential corrections and the improved geometry provided by the geostationary satellites ~~will improve user accuracy~~ to better than 10 m (2drms) in the vertical, which is adequate for aircraft Category I precision approach. The integrity data will improve user safety by flagging GPS satellites that are behaving incorrectly and cannot be corrected. In fact, the WAAS can deliver health warnings to the pilot within 6 s of a GPS satellite malfunction.

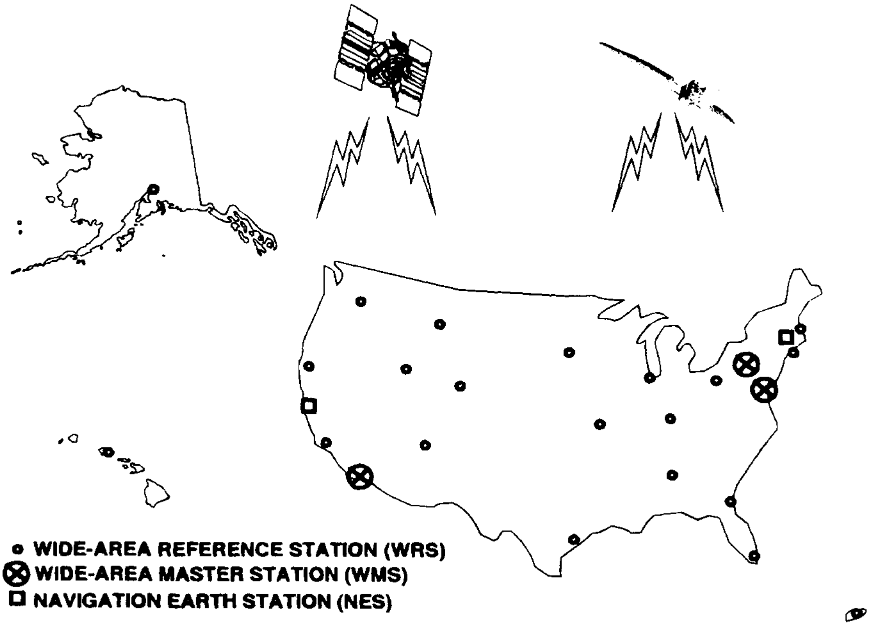

● WIDE-AREA REFERENCE STATION (WRS)
⊗ WIDE-AREA MASTER STATION (WMS)
□ NAVIGATION EARTH STATION (NES)

Fig. 2 Example wide area augmentation system ground network for the United States.

Section II of this chapter details the WAAS signal design, including the link budget, design of the spectrum-spreading codes, and forward error correction. Section III describes the ranging function of the WAAS signal and quantizes the improvement in position-fixing availability made possible by the WAAS ranging sources. It also includes some interesting results that specify the required accuracy of the differential corrections as a function of phase of flight.

Section IV discusses nonprecision approach, where stand-alone GPS provides sufficient accuracy, and the WAAS data only need provide integrity. Section V discusses Category I precision approach, where the WAAS must provide vector corrections to achieve the accuracy requirements. Section VI describes the WAAS data format, which was approved by Working Group 2 of the Radio Technical Committee for Aeronautics (RTCA) in the Spring of 1994, and Sec. VII contains a brief summary. Finally, an appendix briefly describes the estimation of the geostationary satellite ephemeris and clock offsets.

This chapter draws on the work of the RTCA Special Committee 159, which is charged with studying the integrity of GPS and writing performance standards for the airborne equipment.[1,2] In particular, this chapter is based on the efforts of Working Group 2, which is responsible for the WAAS.

Importantly, this chapter does not describe the algorithms that generate the vector corrections, nor does it analyze the accuracy of a positioning system that uses vector corrections. These important topics are covered in Refs. 3 and 4 and the previous chapter.

II. Signal Design

The WAAS signal must achieve the following:

1) Not interfere with the reception of GPS signals by existing GPS receivers or by other GPS receivers that are not designed to receive the WAAS signal

2) Provide an additional ranging measurement

3) Provide the highest possible capacity for integrity data and differential corrections

4) Be received by a modified GPS receiver with minimum complexity

To achieve these goals, the biphase shift keyed (BPSK) signal given in Eq. (1) seems to be the strongest candidate among the many discussed by Working Group 2.

$$s(t) = \sqrt{2C}X(t)D(t)\cos(\omega_L t + \theta) \qquad (1)$$

In this equation, C is the power of the single carrier, and $D(t)$ is the data waveform that modulates the carrier. The carrier phase is θ, and $X(t)$ is a Gold sequence with the GPS chipping rate of 1.023×10^6 chips/s. The carrier frequency is $\omega_L \approx 2\pi\ 1575.42 \times 10^6$ rad/s, which is the same as the GPS L_1 frequency.

A. Link Budget and Noninterference with GPS

As shown in Table 1 and Fig. 3 (see Ref. 5), the received WAAS power through a typical GPS antenna will vary from −161 dBW to −157 dBW depending on the elevation angle. The signal is assumed to be received at a low elevation angle of 5 deg. The table also shows the probability of bit error [$Pr(\epsilon)$] as a function

Table 1 Link power budget for the wide area augmentation system

Received WAAS carrier power	−161 dBW
Receiver antenna gain at 5 deg	−4 dBic
Cable/filter losses	−1 dB
Other implementation losses	−1 dB
Net effective carrier power (C)	−167 dBW
Boltzmann's constant	−228.6 dBW/K Hz
Equivalent noise temperature T_{eq}	+28 dB K
Noise power spectral density N_0	−200.6 dBW/Hz
C/N_0	33.6 dB Hz
Uncommitted margin	4.0 dB
Final C/N_0	29.6 dB Hz
100 b/s	
$E_b/N_0 = C/N_0 - 10 \log_{10} (100)$	9.6 dB
$Pr(\epsilon)$ without forward error correction (FEC)	$\approx 7 \times 10^{-6}$
$Pr(\epsilon)$ with forward error correction (FEC)[6]	$< 10^{-12}$
250 b/s	
$E_b/N_0 = C/N_0 - 10 \log_{10} (250)$	5.6 dB
$Pr(\epsilon)$ without forward error correction (FEC)	$\approx 2.0 \times 10^{-3}$
$Pr(\epsilon)$ with forward error correction (FEC)[6]	$\approx 1.5 \times 10^{-8}$

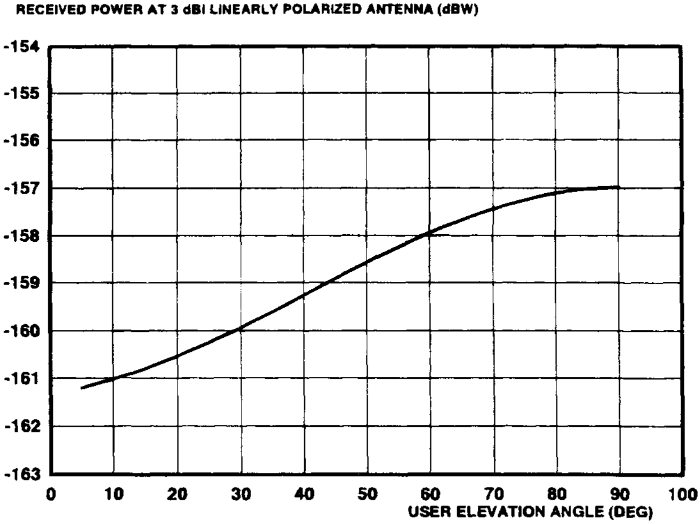


Fig. 3 Expected typical received power levels from the wide area augmentation system.

of data rate. The forward error correction scheme is a $R = 1/2$ convolutional code with a constraint length of 7 and 3 bit soft decisions. This yields C/N_0 ratios of 33.6 dB Hz to 37.6 dB Hz, which in turn corresponds to E_b/N_0 ratios of 5.6–9.6 dB at 250 b/s.

The WAAS signal will not interfere with existing GPS receivers or future GPS receivers that are not designed to receive the WAAS signal. The WAAS power levels are generally weaker than the current and expected power levels for GPS. Even the highest possible WAAS power level is no more than 7 dB above the minimum specified GPS C/A power level of -164 dBW. Additionally, the strongest WAAS power level (-157 dBW) is approximately 20 dB weaker than the ambient noise power in the 2 MHz C/A-code bandwidth (and 30 dB weaker than the ambient noise in the 20 MHz P-code bandwidth).

Because the GPS and WAAS power levels are nearly equal, code division multiple-access (CDMA) can be used to separate the signals and prevent interference. As described in Chapter 3 of the companion volume, the GPS satellites share their L_1 band by using CDMA rather than time division multiple access (TDMA) or frequency division multiple access (FDMA). The GPS satellites use nearly orthogonal codes from a family of 1025 Gold sequences, and the WAAS satellites will broadcast unused Gold sequences from this same family.

Spread-spectrum multiple access is based on the lack of correlation between the different codes that belong to a given set of "signature" sequences. This property is depicted in Fig. 4, which assumes that K signals are sharing a channel, and so the received signal is given by the following:

$$r(t) = s_1(t) + \sum_{k=2}^{K} s_k(t - \tau_k) + n(t)$$

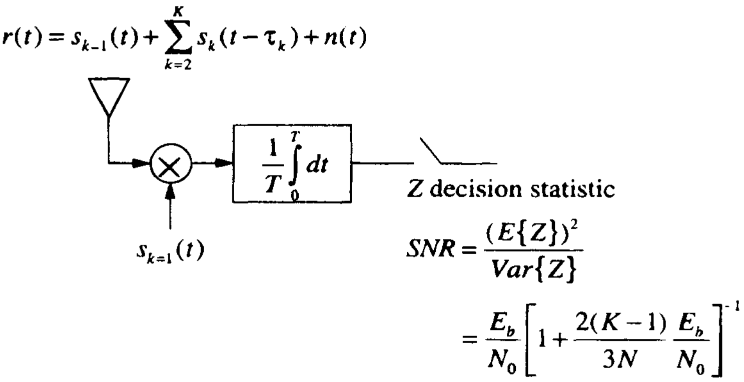

$r(t) = s_{k-1}(t) + \sum_{k=2}^{K} s_k(t-\tau_k) + n(t)$
$\frac{1}{T}\int_0^T dt$
Z decision statistic
$s_{k=1}(t)$
$SNR = \frac{(E\{Z\})^2}{Var\{Z\}}$
$= \frac{E_b}{N_0}\left[1 + \frac{2(K-1)}{3N}\frac{E_b}{N_0}\right]^{-1}$

Fig. 4 The WAAS signal uses code division multiple access to share the L_1 band with GPS.

In this equation, $s_1(t)$ is the signal from the "desired" satellite, and $\{s_k(t - \tau_k)\}_{k=2}^{K}$ are the signals from the competing satellites. The channel also adds additive white Gaussian noise $n(t)$, which has a power spectral density of $N_0/2$. As shown in Fig. 4, the receiver is matched to the signal from the $k = 1$ user and integrates over T seconds.

The signal-to-interference ratio for the receiver shown in Fig. 4 can be computed based on the following assumptions. First, the desired and competing signals are received with equal power. (The equal power assumption can be removed without any real difficulties, but we retain it for simplicity.) Second, the energy in all K signals is given by the following:

$$E_s = \frac{1}{T}\int_0^T s_k^2(t)\mathrm{d}t$$

Third, all signature sequences are comprised of N random binary bits, where these bits are independent of other bits in their own sequence and any other satellite's sequence. These binary bits modulate the carrier using biphase shift keying. Finally, the decision statistic Z is the sampled output of the integrator shown in Fig. 4.

The signal-to-noise (SNR) ratio of that decision statistic will be

$$\mathrm{SNR} = \frac{(E\{Z\})^2}{\mathrm{Var}\{Z\}}$$

$$= \frac{E_s}{N_0}\left[1 + \frac{2(K-1)}{3N}\frac{E_s}{N_0}\right]^{-1}$$

where the full analysis can be found in Ref 7. The signal-to-noise ratio in the absence of multiple-access interference is E_s/N_0; so the $K - 1$ competing users contribute the term inside the square brackets. As shown, the multiple-access interference is manageable provided that N is large compared to K. For GPS, $N = 1023$, and K is typically 8 or 9; thus multiple-access interference is small.

Importantly, if two or three WAAS satellites are added, then K is typically equal to 10 or 12, and the multiple-access interference will remain small.

Of course, GPS does not use random sequences. Rather, the GPS satellites draw their signature sequences from a set of Gold codes, and the WAAS signatures will be unused sequences from the same set. Indeed, any sequence in a Gold set can only take three cross-correlation values with any other sequence in the set, and these three values are shown in Fig. 5.

If the codes achieve their worst-case alignment in time, the magnitude of the cross-correlation is upper bounded by 65/1023, and the processing gain is $20 \log_{10}(1023/65) \approx 24$ dB. This means that the correlation level of the weakest GPS C/A signal will always be at least 17 dB over the strongest WAAS signal. The margin of the P code will be significantly greater, because of the tenfold increase in the chipping rate and the absence of cross-correlation.

A preferred set of codes from the GPS Gold code family has been identified for the WAAS.[8] These codes have the minimum number of shifts where the maximum autocorrelation occurs, and they have the minimum number of shifts where the maximum cross-correlation with the GPS codes occur.

In summary, the WAAS will not interfer with GPS, because the differences in the received power levels are small compared to the spread spectrum processing gains. In fact, WAAS received power levels are generally below that of GPS.

B. Data Capacity

The WAAS signal has enough capacity to carry both the differential corrections and the integrity data required to augment GPS. Recall that the differential corrections will improve user accuracy to better than 10 m (2 drms) in the vertical,

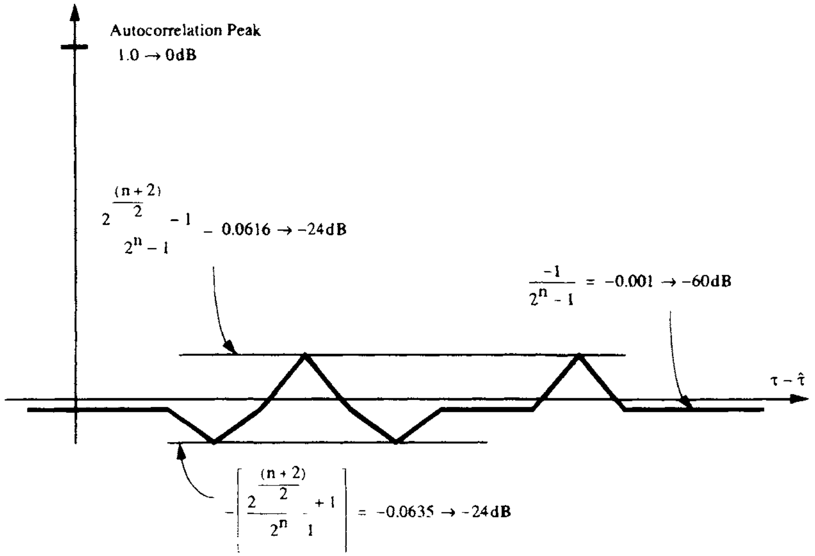


Fig. 5 Worst case cross-correlation between Gold sequences of length 1023.

which is adequate for aircraft Category I precision approach. The integrity data will improve user safety by flagging GPS satellites that are behaving incorrectly and cannot be corrected. For the reasons discussed in Sec. V and VI, these data require approximately 250 b/s of throughput.

Table 1 includes the probability of bit error for data carried by the WAAS as a function of the data rate. As shown, the probability of bit error for 100 b/s data would be sufficiently low even without the use of forward error correction. However, the corrections discussed in Secs. V and VI require 250 b/s; so forward error correction must be used to reach acceptably low bit error probabilities.

The error correction code used in Table 1 is the $R = 1/2$, convolutional code with constraint length 7 because the corresponding transmission rate is 500 symbols per second (sps), which is a multiple of 50 b/s and a submultiple of the C/A-code repetition rate of 1000 Hz. This code introduces a decoding delay of five constraint lengths or 35 databits, which is 0.14 s at 250 b/s. Incidentally, the $Pr(\epsilon)$ values in Table 1 assume three-bit soft decisions, and hard decisions would cost 2 dB.

This convolutional code is attractive, because it is well known and is a standard for satellite communications. Accordingly, decoders are available as integrated circuits for a few tens of dollars. As an alternative, this code can easily be realized in software at the 250 b/s data rate required for the WAAS.

C. Loop Threshold

Although forward error correction lowers the signal-to-noise ratio at which bit errors are made, it does not lower the signal-to-noise ratio at which the phase–lock-loops (PLL) lose lock. The PLL threshold can be estimated from the following equation for the variance of the phase estimate from a Costas Loop:

$$\sigma_\Phi^2 = \frac{B_L N_0}{C}\left[1 + \frac{N_0}{2CT}\right] \text{ rad}^2$$

where B_L is the loop noise bandwidth in Hertz and $1/T$ is the data or symbol rate. If this equation is set to a threshold phase variance of $(0.25)^2$ rad^2, then the SNR threshold is as follows:

$$C/N_0 = 8B_L[1 + \sqrt{1 + 1/(8B_L T)}]$$

or

$$[C/N_0]_{dB} = 10 \log_{10}(C/N_0) \text{ dB Hz}$$

This loop threshold is plotted in Fig. 6. Note that for nominal loop bandwidths (10–15 Hz), there is about a 2 dB difference between the 50 and 500 Hz predetection bandwidth performances.

III. Ranging Function

As described in Sec. II, the WAAS signal will be at the GPS L_1 frequency and will be modulated with a length 1023 Gold sequence with the same chipping rate as the GPS signals. The phase of the Gold code will be synchronized to GPS, and the carrier frequency will be controlled to facilitate carrier aiding. This

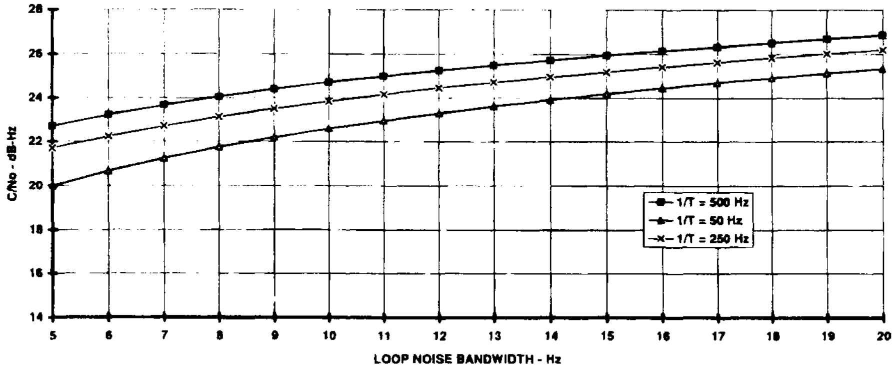


Fig. 6 Approximate tracking threshold for a Phase–lock-loop tracking the WAAS Signal (no dynamics).

code phase control is described in the Appendix to this chapter which is based on the more detailed discussion in Ref. 9.

With such synchronization, the WAAS satellites will provide new range measurements, and improve the *time availability* of position fixing. This improvement is shown in Fig. 7, which is from Ref. 10. Figure 7 shows the availability of a given level of horizontal dilution of precision (HDOP) in the conterminous United States (CONUS) when averaged over 24 hours. It incorporates the effect of possible satellite failures, where GPS satellite failures are based on the model described in Ref. 11. Wide area augmentation system satellite failures are based on the model described by Ref. 12.

Figure 7 shows availability versus HDOP when 0, 1, 2, or 3 geostationary satellites are used with 24 GPS satellites (21 primary plus 3 active spares). The geostationary satellites are located in optimal or nearly optimal locations for covering CONUS. As shown, if 2 geostationary satellites are parked in optimal locations, then an HDOP of 3.2 or better is achieved 99.999% of the time. If no geostationary satellites are used, then an HDOP of 3.2 is only achieved approximately 99.9% of the time. As such, the geostationary satellites reduce the probability of outage by two orders of magnitude.

The results shown in Fig. 7 for two WAAS satellites are probably achievable at reasonable cost. The assumed location of 60 deg W is very close to the planned location for the Inmarsat-3 satellite for the Atlantic Ocean West (55°W). The location at 100°W is near the center of the North American domestic satellite arc, where many system operators have geostationary satellites.

If a barometric altimeter is used, then lower HDOPs can be achieved with an availability of 99.999%.[10] Clearly, if the Glonass satellites, described in Chapter 9 of this volume, are added to the constellation, then lower HDOPs or higher availabilities are also possible.

The horizontal or vertical position error of a user can be bounded based on the corresponding DOP distribution and the pseudorange error standard deviation.[10]

$$\mathrm{HPE}_d \leq 2\mathrm{HDOP}_d\sigma_{PR}$$

$$\mathrm{VPE}_d \leq 2\mathrm{VDOP}_d\sigma_{PR}$$

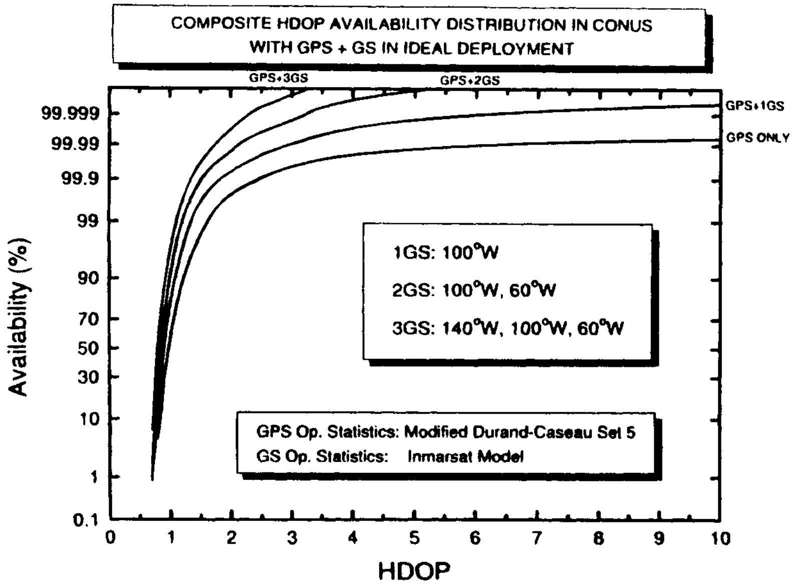


Fig. 7 Availability vs horizontal dilution of precision.[10] This analysis is for ranging WAAS satellites located as shown.

where HPE_d and $HDOP_d$ are the horizontal position error and horizontal dilution of precision not exceeded with probability d. Conversely, these values are exceeded with probability $1 - d$. VPE_d and $VDOP_d$ are similar variables for the vertical dimension. These relationships have been used to plot HPE and VPE vs σ_{PR} in Figs. 8 and 9.

As discussed in the next two sections, the HPE and VPE data shown in Figs. 8 and 9 can be used to derive requirements on the accuracy of the corrections provided by the WAAS. Section IV discusses nonprecision approach, where stand-alone GPS provides enough accuracy, and the WAAS data only need to provide integrity. Section V discusses Category I precision approach, where the WAAS must provide vector corrections to meet the accuracy requirements.

IV. Nonprecision Approach and Error Estimates

Nonprecision approach traditionally uses a baroaltimeter for vertical position information, and a ground-based radio navigation aid for horizontal information. In fact, the horizontal information is typically derived from any of the following (or combinations thereof): a VHF omnidirectional radio (VOR) navigation aid; distance-measuring equipment (DME); Loran-C; or automatic direction finding (ADF) to a nondirectional beacon (NDB). Global Positioning System receivers can be certified for nonprecision approach if they conform to the performance requirements in Technical Standard Order C-129, which was released in 1993. These receivers still derive vertical information from a baroaltimeter, and GPS provides the horizontal information. However, these TSO-C129 receivers are for

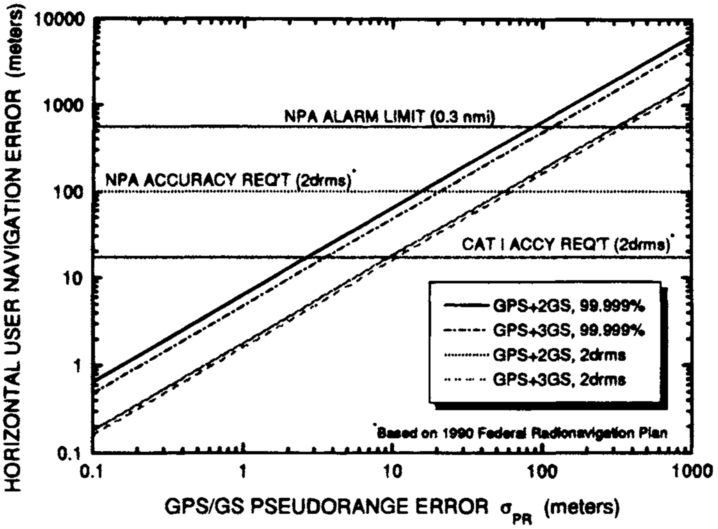


Fig. 8 Horizontal position error vs the standard deviation of the corrected pseudorange.[10]

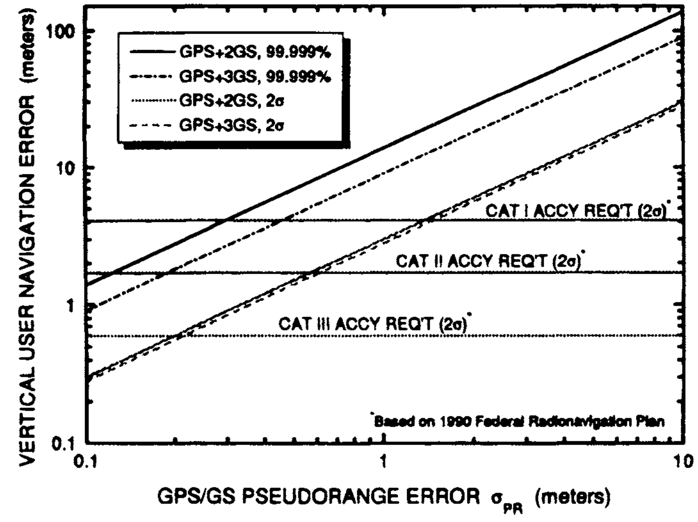


Fig. 9 Vertical position error vs the standard deviation of the corrected pseudorange.[10]

supplemental use only, and some augmentation of GPS, such as the WAAS, is required for GPS to be the primary navigation system.

Strawman specifications for nonprecision approach are superposed onto Fig. 8. As shown, these tentative specifications call for the horizontal error to be less than 100 m 95% of the time, and less than 560 m 99.999% of the time. If GPS is augmented with two WAAS satellites, then the 95 and 99.999% horizontal specifications require that $\sigma_{PR} \leq 60$ ms and $\sigma_{PR} \leq 90$ m, respectively. These derived requirements can be met without the use of differential corrections.

For nonprecision approach, the WAAS does not need to send correction data, but it must warn the user of any GPS malfunctions. Such integrity data could take some very simple forms. For example, it could simply warn the user not to use GPS for navigation, or it could warn the user not to use a specific GPS satellite. These warnings could apply to all phases of flight or they could specify the affected phase. Although these strategies are simple, they will result in conservative alarm thresholds. After all, the ground segment would need to send "don't use" messages for satellites that were unsuitable for use anywhere in the coverage area.

More sophisticated integrity data leave the final "use/don't use" decision to the aircraft receiver, which can account for its own geometry and satellite selection. In this case, the WAAS would send pseudorange error estimates for each satellite. The user would compute its uncorrected position; then it would apply the error estimates as corrections, and compute its corrected position. If the two position fixes differed by more than a certain threshold, then the receiver could discard one satellite at a time until the position fixes were close enough.[13] The receiver would declare the system unavailable only if it could not produce agreement. Consequently, the transmission of error estimates for each satellite would yield a much lower false alarm rate than the transmission of the simpler "don't use" messages described in the last paragraph.

A dense network of reference stations is not required to generate the error estimates for nonprecision approach, because the error estimate need only be moderately accurate. In fact, 10–12 reference station may well provide adequate worldwide coverage.[14] However, denser networks, such as the one shown in Fig. 2, are required to provide sufficient accuracy for precision approach.

V. Precision Approach and Vector Corrections

Unlike nonprecision approach, precision approach is based on a smooth glide path with a constant rate of descent. This glide path, typically 3 deg, passes through a "decision height" at which the pilot must decide whether or not to complete the landing. The pilot must execute a missed approach unless visual references have become available. Precision approach is divided into three categories depending on the decision height and the "runway visual range" (RVR). Category I conditions exist when the decision height is at 200 feet or above, and the RVR is 2400 feet or greater. Category II conditions exist when the decision height is between 100 and 200 feet, and the RVR is 1200 feet or greater. Category III conditions exist when the visibility is poorer, and include conditions with zero visibility. The requirements for Category I, II, and III precision approach are superposed on Fig. 9, which shows vertical position error versus σ_{PR}.

As shown in Fig. 9, the 2 drms (95%) requirements for Category I precision approach require that $\sigma_{PR} \leq 3.0$ m. These requirements can be met using either *scalar* corrections over a local area or *vector* corrections over larger areas. As shown in Fig. 10, local broadcast of scalar corrections can provide Category I capability, but the validity of the corrections decreases with distance. This spatial decorrelation is detailed in Chapter 1 of this volume. Briefly, the error sources that cause the correction to decorrelate spatially are: errors in the GPS ephemeris data, ionospheric refraction, and tropospheric refraction. If a broadcast system is to provide Category I service to a large area (hundreds of miles across), it must deliver *vector* corrections. These corrections are described later in this section and in Sec. VI.

Also shown in Fig. 9, the requirements for Category II or III precision approach demand an extremely accurate pseudorange measurement. These applications require some variety of local area system (very close to the runway) and are beyond the capability of the WAAS.

A. Vector Corrections

Vector corrections carry the following components for each satellite:

Satellite clock offset: This term does not decorrelate spatially, but decorrelates temporally because of selective availability (SA). In fact, the satellite clock corrections must be sent much more frequently than any other component of the vector correction.

Satellite ephemeris: Estimates of all three components of Δr_k are sent, where this vector connects the true location of the kth satellite to the location given by the satellite's navigation message. In component form, these data do not decorrelate spatially and decorrelate very slowly in time. In fact, the update rate for the

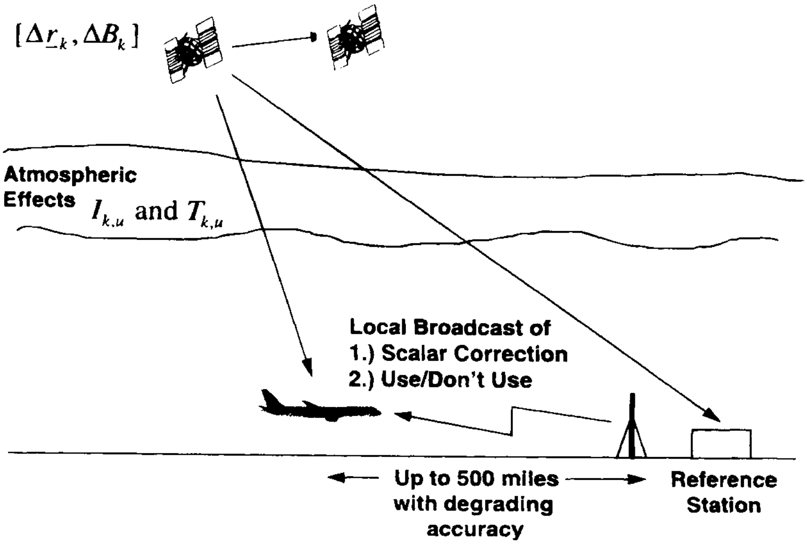

$[\Delta \underline{r}_k, \Delta B_k]$
Atmospheric Effects $I_{k,u}$ and $T_{k,u}$
Local Broadcast of
1.) Scalar Correction
2.) Use/Don't Use
Up to 500 miles with degrading accuracy
Reference Station

Fig. 10 Local area broadcast of scalar corrections.

ephemeris data is not set by concerns over accuracy. Rather, it is set to ensure that the time to first position fix is not too large.

The vector corrections also carry separate data that describe the ionosphere. These data may be a grid of ionospheric samples or a set of coefficients for an orthonormal function description of the ionospheric delay. The data do not decorrelate rapidly in time, and an update interval of 2–5 minutes is probably adequate.

As described in Chapter 1 of this volume, ionospheric corrections lose their validity as the user moves away from the reference station. This concern will dictate the spacing between the reference stations and the amount of ionospheric data sent. In fact, a likely set of reference stations could be located at the Federal Aviation Administration's Air Route Traffic Control Centers (ARTCC). This network is shown in Fig. 2 and inlcudes approximately 24 reference stations.

Vector corrections are designed to serve a large geographical area, and as such they are sometimes called wide area differential GPS (WADGPS) corrections. Wide area differential GPS is well described in the previous chapter, which details this system concept.

B. Precision Approach Integrity

Vector corrections will certainly provide enough accuracy for Category I approach. However, their use raises an interesting integrity issue because the aircraft must now fly on the corrected position fix. In other words, the WAAS has seemingly combined the positioning function with the integrity function, whereas these two functions are traditionally independent. A full analysis of Category I integrity when using the WAAS is beyond the scope of this chapter. However, adequate integrity can be realized with a combination of ground monitoring and receiver autonomous monitoring, as briefly described in the following paragraphs.

1. Ground Monitoring

The set of ground monitors shown in Figs. 1 and 2 help realize the integrity function by setting and broadcasting a "don't use" flag for any satellite. This flag is set only if the ground network cannot develop differential corrections with confidence. For example, if the error acceleration will render the correction invalid before the next update, the master station will set the "don't use" flag for that satellite.

Additionally, the ground monitors could be partitioned into two groups: reference stations and integrity monitors. The latter group would be dedicated to integrity checking and would be independent of the reference stations. The observations made by these integrity monitors are not used to develop the corrections, and if there are any difficulties with the corrected fixes, they simply cause a "don't use" message to be sent.

Alternatively, the ground monitors could all be reference stations, and data from all of the ground stations would be used to form the WAAS corrections. In this case, integrity would be determined by computing the measurement residuals from each reference station. In other words, if a transmitted correction was very different from the correction suggested by a single reference station, a "don't use" message would be sent.

In either case, the density of the WAAS ground stations will be eventually determined by the *integrity* requirements of Category I precision approach. Specifically, they must be dense enough to guarantee that no large error has been introduced into a local area that cannot be detected by the nearest one (or two) monitors.

2. *Receiver Autonomous Integrity Monitoring*

Receivers using the WAAS can confirm the integrity of their position fix by using receiver-autonomous integrity monitoring (RAIM), which is described in the next chapter. For Category I approach, the aircraft applies the differential corrections sent by the WAAS and then determines whether or not the corrected data supplied by the different satellites are *consistent.* This autonomous action protects the aircraft against certain error mechanisms that may not be noticed by a sparse set of ground monitors.

As described in the next chapter, such a consistency check requires that at least five satellites are in view of the aircraft. It also requires that the vertical dilution of precision (VDOP) of all five subsets be acceptably small, where the subsets are formed by deleting one of the satellites in view. Consequently, the time availability of this self-contained integrity monitoring is smaller than the time availability of position fixing. In fact, the low time availability of autonomous fault detection is a main reason that the GPS constellation must be augmented.

As shown in Figs. 11 and 12, the additional pseudoranges provided by the geostationary satellites dramatically improve the availability of autonomous failure detection. Figure 11 is for the 24-satellite GPS constellation with no satellite failures and no additional geostationary ranging sources. Figure 12 is for the same set of GPS satellites but includes four Inmarsat-3 satellites as well as geostationary satellites at the following locations: 100°W and 135°W.

Both figures show the the maximum vertical dilution of precision (VDOP) among the five subsets formed by deleting one satellite at a time. This number is called the "maximum subset VDOP", and the figures show the daily value of maximum subset VDOP, which is not exceeded 99.9% of the time.

As shown in these figures, geostationary satellites significantly increase availability or equivalently decrease the outage probability. Furthermore, if some auxilliary onboard sensor allows the aircraft to coast through short outages, then the probability of outage is further decreased. However, neither figure accounts for satellite failures; and this possibility must be accounted for in a full analysis.[11]

VI. Wide Area Augmentation System Message Format

For each GPS satellite, the WAAS message contains separate corrections for the quickly varying component of the pseudorange error (mostly clock) and the slowly varying component of the pseudorange error (mostly ephemeris). The WAAS message also carries estimates of the ionospheric delay for a grid of locations.

The WAAS message format serves an extremely ambitious goal: provide Category I precision approach accuracy over a continental area using a data rate of

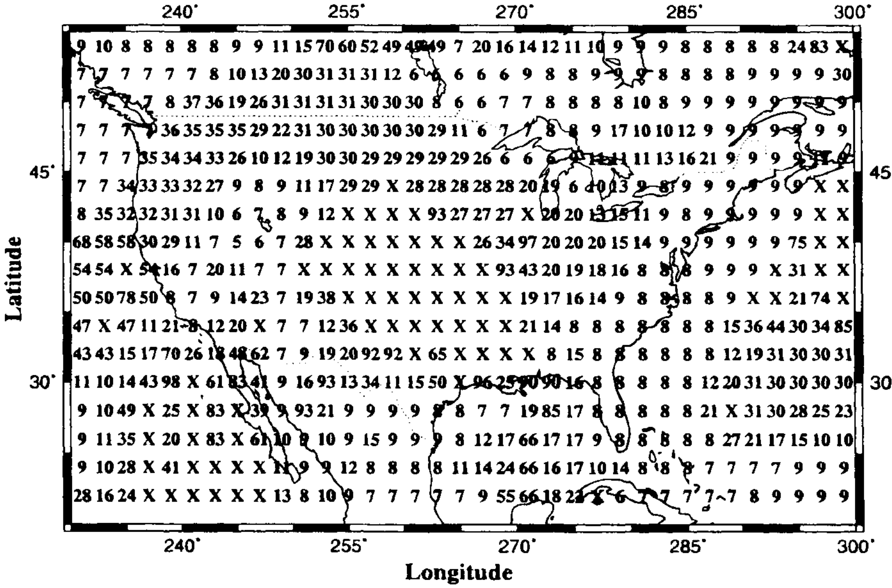


Fig. 11 Subset VDOP for 24 GPS Satellites. The subset VDOP is the maximum VDOP obtained when one satellite is deleted from the set of satellites in view. The value displayed in the figure is the daily value of subset VDOP, which is not exceeded 99.9% of the time. All 24 GPS satellites are assumed to be healthy.

only 250 b/s. To do this, the message stream must carry corrections for all 24 GPS satellites and any GNSS (Global Navigation Satellite System) satellites in orbit. In contrast, local area DGPS datalinks typically use 100 b/s or more to provide a similar capability to ranges of 200–300 km. Moreover, local links need only carry corrections for the 6–12 satellites in view of the single reference station.

The WAAS message format is extremely efficient for the following reasons:

Fast corrections: These messages carry the quickly varying component of the pseudorange errors for each satellite (mostly clock error). They must be sent much more frequently than any other message (every 6–10 s), but they do not decorrelate spatially. Consequently, a single, very short message suffices for the entire footprint of the geostationary WAAS satellite. The GPS error components, which do decorrelate spatially, are corrected by separate messages that need not be sent frequently.

No rate corrections: Unlike most local area DGPS data formats, the WAAS carries no rate corrections for the quickly varying component of the satellite error. As shown in Ref. 16, it is more efficient for the user receiver to estimate the rate by differencing the most recent fast corrections.

Masks: A "mask" is used to designate which satellite belongs to which slot in the fast correction messages. A mask is used to assign slots, because the WAAS data format must be able to send corrections efficiently for all GNSS satellites in orbit. At times, all 24 satellites in the GPS primary 21 (+ 3) satellite constellation can be seen by a combination of users in the footprint of a geostation-

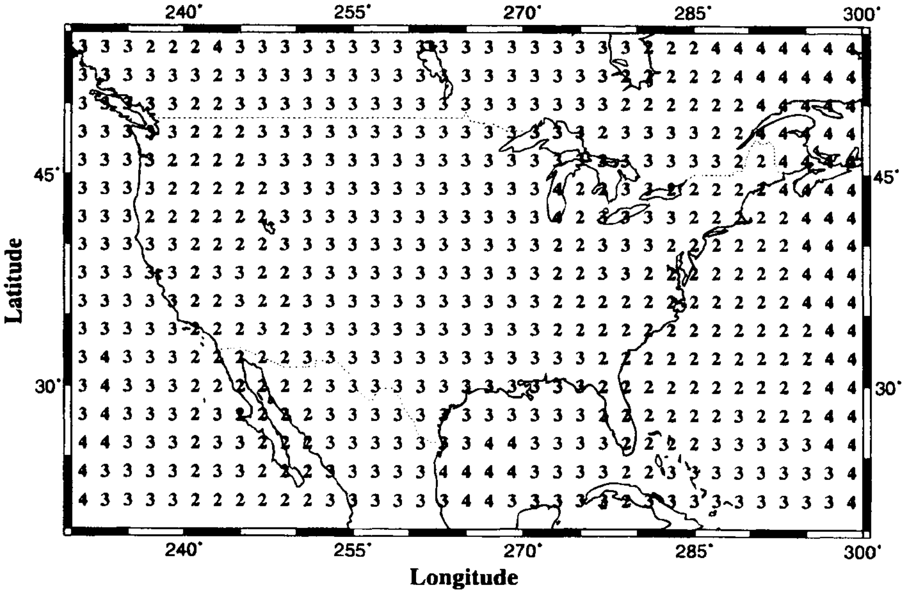


Fig. 12 Subset VDOP for 24 GPS satellites plus 4 Inmarsat-3 satellites and 2 additional geostationary satellites. The subset VDOP is the maximum VDOP obtained when one satellite is deleted from the set of satellites in view. The value displayed in the figure is the daily value of subset VDOP, which is not exceeded 99.9% of the time. All 24 GPS satellites are assumed to be healthy.

ary satellite.[17] A similar ionospheric mask is used to associate each slot in the ionospheric correction message with a geographic location.

Geostationary navigation message: In contrast to the GPS satellites, the WAAS satellites are geostationary. Hence, they have much lower accelerations than the GPS satellites, and their ephemeris need not be updated as frequently.

Parity: As described shortly, the WAAS message uses a much more powerful algorithm for detecting transmission errors than GPS or most local area DGPS datalinks. WAAS error detection only adds 24 parity bits to 226 databits, whereas most local area DGPS data formats add 6 bits to 24 databits. As such, the overhead for parity is reduced from 6/30 = 0.20 to 24/250 ≈ 0.1.

The basic WAAS message is shown in Fig. 13. As shown, all WAAS messages are 250 bits in length. At the data rate of 250 b/s, the duration of a WAAS

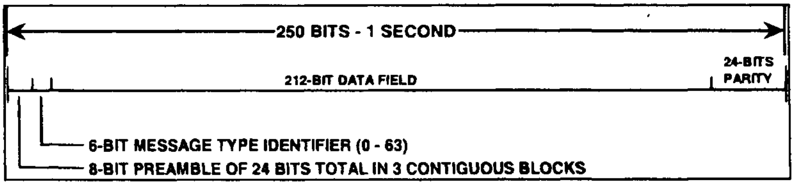

250 BITS - 1 SECOND
212-BIT DATA FIELD
24-BITS PARITY
6-BIT MESSAGE TYPE IDENTIFIER (0 - 63)
8-BIT PREAMBLE OF 24 BITS TOTAL IN 3 CONTIGUOUS BLOCKS

Fig. 13 Basic WAAS datablock format.

message is 1 s, and the start of the message block is synchronous with the 6 s GNSS time epoch.

Each block consists of the following:

1) 8-bit (distributed) preamble
2) 6-bit message type
3) 212-bit datafield
4) 24-bit CRC (cyclic redundancy check) parity

The preamble is a 24-bit unique word, distributed over three successive blocks. An 8-bit preamble is adequate because the WAAS message is 1 s in duration and remains in synchronism with the GPS time epoch of 6 s. The message type field is 6-bits long, which allows for 64 different messages. The currently defined message types are summarized in Table 2.

The following subsections describe the parity field, and then they briefly discuss the type 1, 2, 3, 9, 24, 25, and 26 messages. All message types are detailed in Ref. 5.

A. Parity Algorithm

The WAAS uses a much stronger parity algorithm than the extended Hamming code used in the GPS navigation message (and typically used for the transmission of differential corrections). The GPS parity algorithm uses the extended Hamming code to add 6 parity bits to a field of 26 databits. It achieves a minimum distance of 4, which means that any combination of 3 or fewer bit errors can be detected. However, if an interference or noise burst coincides with one of the 30-bit words, and the received data are completely garbled, then the probability that

Table 2 Wide area augmentation system

Type	Contents
0	Don't use GEO for anything (for testing)
1	Pseudorandom noise (PRN) mask assignments, set up to 50–212 bits
2	Fast pseudorange error estimates
3–8	Reserved for future messages
9	GEO navigation message (*x,y,z,* time, etc.)
10–11	Reserved for future messages
12	WAAS network universal coordinated time (UTC) offset parameters
13–16	Reserved for future messages
17	GEO satellite almanacs
18	Ionospheric pierce point mask 1
19	Ionospheric pierce point mask 2
20	Ionospheric pierce point mask 3
21	Ionospheric pierce point mask 4
22	Ionospheric pierce point mask 5
23	User differential range errors (UDRE) zone radii and weights
24	Mixed fast/long-term satellite error estimates
25	Long-term satellite error estimate
26	Ionospheric delay error estimate
27–63	Reserved for future messages

the algorithm fails to detect the error burst is $p_{fd} = 2^{-6} = 1.56 \times 10^{-2}$ (see Ref. 6). This probability is too large for a system designed to make GPS fail-safe.

To reduce the probability of failing to detect a bit error, the parity scheme in Fig. 13 uses 24 bits and thus reduces the probability of failing to detect an interference burst to $2^{-20} = 9.54 \times 10^{-7}$. As described earlier, the overhead for parity is reduced from 6/30 = 0.20 to 24/250 ≈ 0.096. The WAAS error detection algorithm does increase the data latency of the integrity data, because 250 bits must be collected by the receiver for a parity check instead of only 30. However, this increases transmission time by only 0.88 s at the WAAS datarate of 250 b/s.

B. Message Type 2 Fast Corrections and User Differential Range Errors

The format for type 2 messages is shown in Fig. 14. These messages carry the quickly varying component of the pseudorange errors for each satellite (mostly clock errors). Each message delivers 16-bit datablocks for 13 satellites. The 16-bit datablock consists of a 12-bit correction and 4 bits of user differential error indication (UDREI). Each correction shall have 0.25-m/resolution and a range of ± 255.5 m. A value of +255.5 (011111111111) indicates that the satellite is not currently observed by the ground network. A −256 (100000000000) is a "don't use" message for that satellite.

As shown, message type 2 also contains a two-bit block identification that indicates those satellites to which the corrections apply. Block 0 contains the corrections for the first 13 satellites designated by the mask, Block 1 contains the fast corrections for satellites 14–26, and so forth. If six or fewer corrections remain at any time, they should be contained in a type 24 mixed corrections message. The last half of message type 24 is reserved for the long-term corrections described below.

In the absence of an emergency, the fast corrections could be sent as infrequently as every 10 s. However, if a satellite fails suddenly, then a fast correction message that carries the emergency "don't use" flag can be completed within 6 s.

C. Type 25: Long-Term Satellite Error Corrections Message

This message carries the slowly varying clock and ephemeris components of the pseudorange errors for each satellite. Its format is shown in Fig. 15. As shown, the datafield begins with a velocity code. If this bit is set to 0, then the message will contain the slowly varying clock and ephemeris errors for four satellites. If it is set to 1, then the message contains error *and* error rate estimates, but only for two satellites.

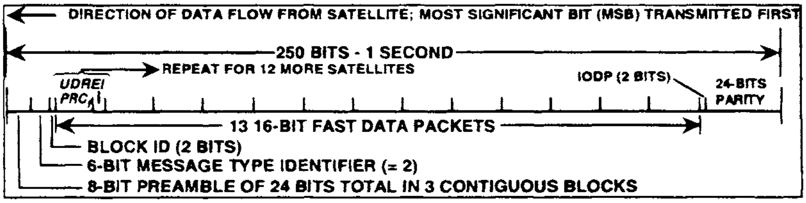

DIRECTION OF DATA FLOW FROM SATELLITE; MOST SIGNIFICANT BIT (MSB) TRANSMITTED FIRST
250 BITS - 1 SECOND
REPEAT FOR 12 MORE SATELLITES
UDREI
PRC
IODP (2 BITS)
24-BITS PARITY
13 16-BIT FAST DATA PACKETS
BLOCK ID (2 BITS)
6-BIT MESSAGE TYPE IDENTIFIER (= 2)
8-BIT PREAMBLE OF 24 BITS TOTAL IN 3 CONTIGUOUS BLOCKS

Fig. 14 Message type 2—fast corrections and UDRE.

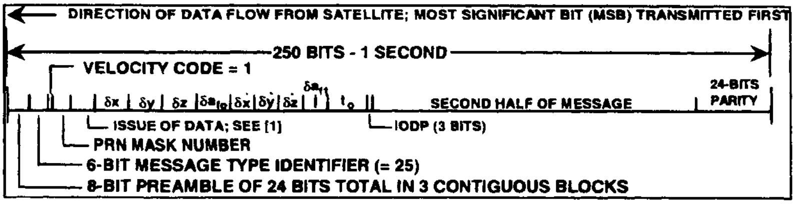


Fig. 15 Message type 25—long-term corrections with velocity code set to 1.

The details of the long-term error correction message are given in Table 3. The ephemeris and ephemeris rate corrections are given in Earth-centered, Earth-fixed (ECEF) coordinates as (Δx_k, Δy_k, Δz_k) and ($\Delta\dot{x}_k$, $\Delta\dot{y}_k$, $\Delta\dot{z}_k$). In Table 3, the clock and clock rate corrections are denoted a_{f0} and a_{f1} respectively. The velocity corrections should be used to predict future ephemeris and clock values using t_0 as the reference time.

D. Type 26: Ionospheric Delay Error Corrections Message

These messages give the WAAS estimates of vertical delay caused by the ionosphere and the 99.5% accuracy of this estimate. The vertical delays are for ionospheric pierce points (IPP), which are specified by the ionospheric mask messages (type 18, 19, 20, 21, and 22). The mask messages specify which of 929 possible IPP locations are being used by the type 26 messages.[5] For example, approximately 80 IPP are required to provide Category I accuracy throughout the National Airspace System (NAS).

Table 3 Long-term error correction message (half message when velocity code is set to 1)

Parameter	Number of bits	Scale factor LSB	Effective range	Units
For each of 1 or 2 satellites	106			
Velocity code	1			discrete
PRN mask number	6	1	50	discrete
Issue of data	8	1	255	discrete
Δx_k (ECEF)	9	0.5	±128	m
Δy_k (ECEF)	9	0.5	±128	m
Δz_k (ECEF)	9	0.5	±128	m
a_{f0}	10	2^{-30}	$\pm 2^{-21}$	s
$\Delta\dot{x}_k$ (ECEF)	7	2^{-10}	±0.0625	m/s
$\Delta\dot{y}_k$ (ECEF)	7	2^{-10}	±0.0625	m/s
$\Delta\dot{z}_k$ (ECEF)	7	2^{-10}	±0.0625	m/s
a_{f1}	7	2^{-38}	$\pm 2^{-32}$	s/s
Time of applicability t_0	16	16	604,784	s
IODP	3	1	0–7	discrete

LSB denotes least significant bit; ECEF denotes Earth-centered Earth-fixed; and IODP denotes issue of data PRN where PRN denotes pseudorange noise code.

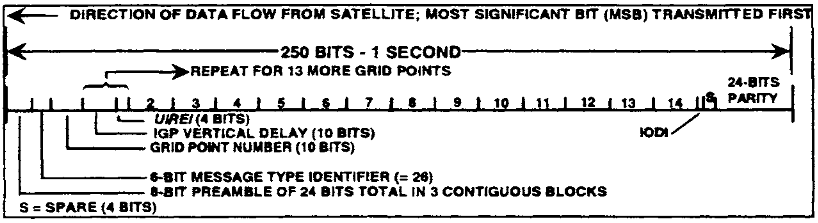

DIRECTION OF DATA FLOW FROM SATELLITE; MOST SIGNIFICANT BIT (MSB) TRANSMITTED FIRST
250 BITS - 1 SECOND
REPEAT FOR 13 MORE GRID POINTS
24-BITS PARITY
2
3
4
5
6
7
8
9
10
11
12
13
14
S
UIREI (4 BITS)
IGP VERTICAL DELAY (10 BITS)
GRID POINT NUMBER (10 BITS)
IODI
6-BIT MESSAGE TYPE IDENTIFIER (= 26)
8-BIT PREAMBLE OF 24 BITS TOTAL IN 3 CONTIGUOUS BLOCKS
S = SPARE (4 BITS)

Fig. 16 Message type 26—ionospheric corrections.

A type 26 message is shown in Fig. 16 and described in Table 4. It carries 13 vertical delay estimates, where each of these is 10 bits and covers a range of 64 m with a resolution of 0.125 m. It also carries the corresponding vertical delay accuracy indicators, where each of these is 4-bits long. The type 26 message also includes 10 bits to identify where the pierce point falls within the current ionospheric mask.

E. Type 9: WAAS Satellite Navigation Message

The navigation message for the geostationary WAAS satellite is shown in Fig. 17, and its contents are detailed in Table 5. The location and velocity of the satellite is described using ECEF coordinates, because transmission of ECEF coordinates is more efficient for a geostationary satellite than transmission of Keplerian elements.

The WAAS navigation message also includes the time of week (TOW) and week number, which represent the time of transmission at the start of the block carrying the message. The TOW can be used to initialize receiver timing during signal acquisition. The TOW is also the reference time for applying the velocity terms to predict position. This message also includes an accuracy exponent to estimate the accuracy of the position information.

F. Applied Range Accuracy Evaluation

The data in the various WAAS messages can be combined to estimate the 99.5% accuracy of the corrected pseudorange. The applied range accuracy (ARA)

Table 4 Type 26 ionospheric delay model parameters

Parameter	Number of bits	Scale factor LSB	Effective range	Units
Pierce point number	10	——	——	discrete
IODI	2	1	0–3	discrete
For each of 13 more grid points	14	——	——	——
Vertical delay estimate	10	1	929	discrete
User ionospheric vertical	4	1	0–15	discrete
Accuracy Indicator (UIREI)	——	——	——	——

IODI denotes issue of data ionosphere; UIREI denotes user ionospheric range error indicator.

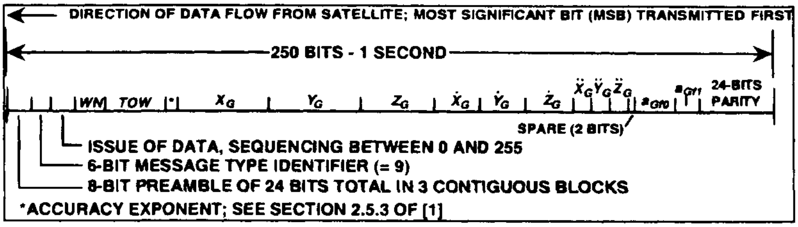

DIRECTION OF DATA FLOW FROM SATELLITE; MOST SIGNIFICANT BIT (MSB) TRANSMITTED FIRST
250 BITS - 1 SECOND
WN
TOW
*
X_G
Y_G
Z_G
$\dot{X}_G$
$\dot{Y}_G$
$\dot{Z}_G$
$\ddot{X}_G$
$\ddot{Y}_G$
$\ddot{Z}_G$
a_{Gf0}
a_{Gf1}
24-BITS PARITY
SPARE (2 BITS)
ISSUE OF DATA, SEQUENCING BETWEEN 0 AND 255
6-BIT MESSAGE TYPE IDENTIFIER (= 9)
8-BIT PREAMBLE OF 24 BITS TOTAL IN 3 CONTIGUOUS BLOCKS
*ACCURACY EXPONENT; SEE SECTION 2.5.3 OF [1]

Fig. 17 Message type 9—Geostationary satellite navigation message.

for satellite i is as follows:

$$\mathrm{ARA}_i = \sqrt{\mathrm{UDRE}_i^2 + F_i^2\,\mathrm{UIREI}_i^2}\ \mathrm{m}$$

The UDRE_i estimate the accuracy of the combined fast and slow corrections and are broadcast in message type 2. They are for the worst location in the coverage region. The UIREI_i estimate the 99.5% accuracy of the ionospheric delay estimates. This estimate must be multiplied by the satellite's obliquity factor F_i, because the UIREI_i value is the error for the vertical ionospheric delay at the pierce point.

VII. Summary

This chapter describes a signal design for a ranging WAAS. In particular, it demonstrate that the proposed WAAS signal can be received by slightly modified GPS receivers without interfering with nonparticipating receivers. The WAAS signal does not interfere, because the received WAAS signal has approximately the same power as GPS signals, and CDMA is used to share the L_1 channel.

The WAAS signal can be controlled to add a valuable range measurement to the navigation solution of airborne receivers. This augmentation significantly improves the availability of position fixing, as well as autonomous fault detection.

Table 5 Wide area augmentation system satellite navigation message parameters

Parameter	Number of bits	Scale factor LSB	Effective range	Units
X (ECEF)	32	0.02	±42,949673	m
Y (ECEF)	32	0.02	±42,949673	m
Z (ECEF)	32	0.02	±42,949673	m
$\dot{X}$ (ECEF)	18	0.0002	±26.2144	m/s
$\dot{Y}$ (ECEF)	18	0.0002	±26.2144	m/s
$\dot{Z}$ (ECEF)	18	0.0002	±26.2144	m/s
TOW	20	1	604,799	s
Week number	10	1	1023	weeks
Accuracy	4			
Issue of data	8	1	255	discrete
Spare	20			

If forward error correction is used, then the WAAS signal has a data capacity of 250 b/s, which is enough to carry the vector corrections required to make possible category I precision approach. In fact, the WAAS uses an extremely efficient data format to carry the required information over the available capacity. This format separates the fast corrections (mostly clock) for each satellite from the slow corrections (mostly ephemeris). It also has separate messages for the ionospheric corrections.

Appendix: Geostationary Satellite Ephemeris Estimation and Code-Phase Control

This appendix briefly describes the WAAS satellite ephemeris estimation process and how the WAAS code phase is controlled. It is based on Ref. 9, where a more detailed explanation is provided. In this appendix, the ephemeris $\Delta \underline{r}_{\text{geo}}$ and clock offset (ΔB_{geo}) of the WAAS satellite are the estimanda of interest. These are estimated using WAAS pseudorange residuals measured at the remote reference stations and forwarded to the master station. The reference station clock offsets Δb_m are "nuisance parameters" and must be estimated using GPS common-view time transfer.

This appendix begins by discussing the use of GPS common-view time transfer to estimate the reference station clock offsets. Then it discusses the determination of the WAAS ephemeris and clock offset when the WAAS is a processing transponder with the clock onboard the satellite. Finally, it discusses WAAS ephemeris estimation and code phase control for a bent-pipe transponder where the clock is on the ground.

For simplicity, this appendix assumes that: the WMS is colocated with the Mth reference station; the WAAS master clock is also located at the WMS; and the uplink to the geostationary satellite is at this same site.

Reference Station Clock Synchronization

The WMS uses common-view time transfer to estimate the difference between the reference station clocks and the master clock located at the WMS. More specifically, it uses the following GPS pseudorange residuals from the M reference stations to estimate the clock offsets:

$$\{(\Delta_{k,m} = \mathbf{\Delta \underline{r}_k} \cdot \mathbf{\underline{1}_{k,m}} + \Delta b_m - \Delta B_k + \nu_{k,m})_{k=1}^{K_m}]_{m=1}^{M} \tag{A1}$$

In Eq. (A1) $\Delta \underline{r}_k$ is the vector that connects the true location of the kth satellite to its location according to the navigation message. In Eq. (A1), $\underline{1}_{k,m}$ denotes the unit vector from the kth satellite to the mth reference station. Additionally, Δb_m and ΔB_k are the offsets in the reference station and satellite clocks. The measurement noise is given by $\nu_{k,m}$; K_m is the number of satellites in view of the mth reference station; and M is the total number of reference stations.

These GPS pseudoranges are called "iono free", because dual-frequency measurements have been used to estimate the ionospheric delay accurately, and this delay is removed from the measured pseudoranges. In addition, they are called "tropo free", because local meteorological measurements have been used to estimate the tropospheric delay accurately, and this delay is also removed.

The master station uses common-view time transfer to estimate the difference between the reference station clocks and the master clock located at the WMS. An estimate of this difference for the *m*th reference station is given by the following:

$$\begin{aligned}\Delta\hat{b}_{m,M} &= \frac{1}{K_{M,m}} \sum_{k=1}^{K_{M,m}} (\Delta_{k,M} - \Delta_{k,m}) \\ &\approx \Delta b_M - \Delta b_m\end{aligned} \tag{A2}$$

where $K_{M,m}$ is the number of satellites in common view of the WMS and the *m*th reference station. This family of estimated clock offsets $(\Delta\hat{b}_{m,M})_{m=1}^{M-1}$ is very important. It is used to eliminate the clock differences in the observations from the reference stations.

It is hoped that the master clock is kept in close synchronism with GPS time. This can be done by averaging GPS observations at the master station to steer an atomic clock or ensemble of atomic clocks. Alternatively, it could be done if there were a more direct connection from the master clock to GPS time as kept by the U.S. Naval Observatory. For example, if a reference station is located at USNO, then it could serve as the timing reference for the entire system.

Processing Transponder

The master station will also estimate the ephemeris of the geostationary WAAS satellite from the pseudorange residuals provided by the reference stations. If the WAAS satellite is a processing transponder with the clock onboard the satellite, then the WAAS residuals from the reference stations are the following:

$$(\Delta_{\text{geo},m} = \Delta\underline{r}_{\text{geo}} \cdot \underline{1}_{\text{geo},m} + \Delta b_M - \Delta B_{\text{geo}} + \nu_{\text{geo},m})_{m=1}^{M} \tag{A3}$$

where ΔB_{geo} is the WAAS satellite clock offset relative to the data in the WAAS navigation message, and $\underline{r}_{\text{geo}}$ is the error in the broadcast WAAS ephemeris. This equation assumes that the reference station clock offsets have been removed. These measurements are "tropo-free," because local meteorological measurements have been used to estimate the tropospheric delay accurately, and these delay estimates have been subtracted from the measured residuals.

The ionospheric delay is removed from the WAAS measurements in one of two ways. First, the WAAS downlink will include a signal at C-band. As such, dual-frequency measurements can be used to estimate the ionospheric delay. However, the C-band signal will be considerably weaker than the L_1-band signal; thus a dish antenna will be required. Alternatively, the GPS measurements at the reference station can be used to build an accurate model of the local ionosphere, and that model can be used to estimate the WAAS ionosphere.

These WAAS residuals from all *M* reference stations form the following linear system of equations:

$$\begin{aligned}\begin{bmatrix}\Delta_{\text{geo},1}\\ \Delta_{\text{geo},2}\\ \vdots\\ \Delta_{\text{geo},M}\end{bmatrix} &= \begin{bmatrix}\underline{1}_{\text{geo},1} & 1\\ \underline{1}_{\text{geo},2} & 1\\ \vdots & \vdots\\ \underline{1}_{\text{geo},M} & 1\end{bmatrix} \begin{bmatrix}\Delta\underline{r}_{\text{geo}}\\ \Delta b_M - \Delta B_{\text{geo}}\end{bmatrix} + \underline{\nu} \\ &= G_{\text{geo}}\Delta\underline{r}_{\text{geo}} + \underline{\nu}\end{aligned} \tag{A4}$$

This system of equations is solved by inverting or "pseudo-inverting" the design matrix G_{geo}, which depends solely on the geometry of the WAAS satellite relative to the reference stations. The resulting estimates of $\Delta \underline{r}_{\text{geo}}$ and $\Delta b_{\text{M}} - \Delta B_{\text{geo}}$ will have minimum mean square error. The operational master site will use a Kalman filter based on these equations to estimate these states as well as their rates as a function of time. The WMS then includes these data in the type 9 message described earlier.

Bent-Pipe Transponder

If the WAAS satellite is a bent-pipe transponder with the clock on the ground, then the WAAS residuals from the reference stations are as follows:

$$\{\Delta_{\text{geo},m} = \Delta \underline{r}_{\text{geo}} \cdot 1_{\text{geo},m} + \Delta b_M + \nu_{\text{geo},m} \\ + (|r_{M,\text{geo}}| + I_{M,\text{geo}} + T_{M,\text{geo}} - B_{\text{geo}} - B_{\text{uplink}} - B_{\text{down}} + \Delta\tau)\}_{m=1}^{M} \qquad \text{(A5)}$$

The first part of this Eq. (A5) is the normal one-way GPS pseudorange residual. It is the residual after the following terms have been removed: the nominal one-way range from the WAAS satellite to the *m*th reference station; the clock offset between the *m*th reference station clock and the master clock; the downlink ionosphere; and the downlink troposphere.

The second line in Eq. (A5) above contains the terms unique to a bent-pipe transponder and which require some special attention. For Inmarsat-3, the WAAS signal is uplinked from the ground at *C*-band, translated from *C*-band to *L*-band in the satellite transponder, and broadcast to the users at *L*-band. Consequently, the uplink range ($|\underline{r}_{M,\text{geo}}|$) from the control site to the WAAS satellite is included in the pseudorange. The uplink ionosphere and troposphere $I_{M,\text{geo}}$ and $T_{M,\text{geo}}$ are also included, as is the transponder delay B_{geo}. The uplink and downlink hardware delays appear as B_{uplink} and B_{down} respectively. Finally, $\Delta\tau$ is the control, which the WMS exerts over the WAAS code phase.

Most importantly, all of the terms unique to the bent-pipe transponder are common-mode. In other words, they are the same for all M reference stations. Consequently, the system of equations used above for ephemeris estimation can also be used with a bent-pipe transponder. Then the WMS will actually control the transmitted code phase $\Delta\tau$ so that the WAAS "clock" state is nulled. This control policy yields a residual at the user that is compatible with the GPS code phase residuals after correction.

If the WAAS satellite is a bent-pipe transponder, then the master station must also control the frequency of the WAAS uplink signal. This control must yield a signal that can be used by a GPS receiver to "carrier-aid" their code tracking loops. Such control can be achieved by using the dual-frequency WAAS observables to remove the ionospheric delay and by maintaining coherency between the carrier and the code at the master site.

References

[1]RTCA, "Minimum Aviation System Performance Standards—GPS Integrity Channel Working Group," RTCA/DO-202, 1988.

[2]RTCA, "Minimum Operational Performance Standards (MOPS) for Airborne Supplemental Navigation Equipment Using GPS," RTCA 204-91/SC159-293. (These MOPS are modified by Technical Standard Order TSO-C129, released Dec. 10, 1992.)

[3]Kee, C., Parkinson, B. W., and Axelrad, P., "Wide Area Differential GPS," *Navigation,* Vol. 38, No. 2, 1991, pp. 123–143.

[4]Kee, C., and Parkinson, B. W., "High Accuracy GPS Positioning in the Continent: Wide Area Differential GPS," *Proceedings of the Second International Symposium on Differential Satellite Navigation Systems (DNSN-93),* (Amsterdam), March 1993.

[5]RTCA Special Committee 159, Working Group 2, "Wide Area Augmentation System Signal Specification," March 1994.

[6]Peterson, W. W., and Weldon, E. J., "Error Correcting Codes," MIT Press, 1972.

[7]Sarwate, D. V., and Pursley, M. B., "Crosscorrelation Properties of Pseudorandom and Related Sequences," *Proceedings of the IEEE,* Vol. 68, 1980, pp. 593–619.

[8]Nagle, J., Van Dierendonck, A. J., and Hua, Q. D., "Inmarsat-3 Navigation Signal C/A Code Selection and Interference Analysis," *Proceedings of the Annual Technical Meeting of the Institute of Navigation,* (San Diego, CA), Jan. 1992, to appear in *Navigation.*

[9]Van Dierendonck, A. J., and Elrod, B. D., "Ranging Signal Control and Ephemeris/Time Determination for Geostationary Navigation Payloads," *Proceedings of the 1994 Institute of Navigation Technical Meeting,* (San Diego, CA), Institute of Navigation, Washington, DC, Jan. 1994, pp. 393–402.

[10]Phlong, W. S., and Elrod, B. D., "Availability Characteristics of GPS and Augmentation Alternatives," *Proceedings of the 1993 National Technical Meeting of the Institute of Navigation,* (San Francisco), Institute of Navigation, Washington, DC, Jan. 1993.

[11]Durand, J. M., Michal, T., and Bouchard, J., "GPS Availability, Part I," *Navigation,* Vol. 37, No. 2, 1990, pp. 123–140.

[12]Kinal, G. V., "A Note on Satellite Reliability as Applied to the Inmarsat-3 Navigation Payload," Paper presented to Working Group 2 of RTCA SC-159, Jan. 1992.

[13]Virball, V. G., Enge, P. K., Michalson, W., and Levin, P., "A Fault Detection and Exclusion Algorithm to be Used with the GPS Integrity Channel," Record of the IEEE Position Location and Navigation Symposium, Las Vegas, April 11–15, 1994.

[14]Enge, P., Levin, P., Kalafus, R., McBurney, P., Daly, P., and Nagle, J., "Architecture for a Civil Integrity Network Using Inmarsat," *Proceedings of the Third International Technical Meeting of the Satellite Division of the Institute of Navigation,* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 1990, pp. 287–296.

[15]Ananda, M., Munjal, P., Siegal, B., Sung, R., and Woo, K. T., "Proposed GPS Integrity and Navigation Payload on DSCS," Record of the 1993 IEEE Military Communications Conference, Boston, Oct. 11–14, 1993.

[16]Hegarty, C., "Optimal Differential GPS for a Data Rate Constrained Broadcast Channel," *Proceedings of the ION GPS-93, Sixth International Technical Meeting of the Satellite Division of the Institute of Navigation,* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 22–24, 1993, pp. 1527–1535.

[17]Van Dyke, K., "Satellite Visibility Analysis," Paper presented to RTCA SC 159 Working Group 2, RTCA Paper 596-92/SC 159-389, Aug. 18, 1992.

Chapter 5

Receiver Autonomous Integrity Monitoring

R. Grover Brown*
Iowa State University, Ames, Iowa 50011

I. History, Overview, and Definitions

NAVIGATION system integrity refers to the ability of the system to provide timely warnings to users when the system should not be used for navigation.[1] The basic Global Positioning System (GPS) as described in Part I provides integrity information to the user via the navigation message, but this may not be timely enough for some applications, especially in civil aviation. Therefore, additional means of providing integrity are being planned, and two different approaches are being considered. One of these is the receiver autonomous method, now referred to simply as RAIM (receiver autonomous integrity monitoring). A variety of RAIM schemes have been proposed, and all are based on some kind of self-consistency check among the available measurements. Of course, there must be some redundancy of information in order for RAIM to be effective. The other approach to providing an independent assurance of integrity is to have a network of ground monitoring stations whose primary purpose is to monitor the health of the GPS satellites. Appropriate integrity information is then transmitted to users via a radio link of some sort. This is referred to as the GPS integrity channel (GIC). We will be primarily concerned with RAIM in this section. A discussion of GIC is presented in Chapter 24, this volume.

Work on RAIM began in earnest in the latter half of the 1980s, and a wealth of papers have appeared in the navigation literature since then. We do not attempt to give a complete bibliography here. One has only to browse through the proceedings of the meetings of the Institute of Navigation (ION) beginning in 1986 to get a reasonably complete history of the evolution of RAIM. We should note that early papers did not use the term RAIM. The term "self-contained" was more common then. The acronym "RAIM" was first suggested in 1987 by R.M. Kalafus, and it has been used almost universally ever since.[2] Two early papers presented at an ION meeting in 1986 are of special interest, because they

*Distinguished Professor Emeritus, Electrical and Computer Engineering Department.

illustrate two different approaches to RAIM.[3,4] The Lee paper[3] is a good example of what is now referred to as a snapshot scheme. With this method, only current redundant measurements are used in the self-consistency check. On the other hand, the Brown and Hwang paper[4] presents a scheme where both past and present measurements, along with a priori assumptions with regard to vehicle motion, are used in the RAIM decision. Such schemes are loosely referred to as averaging or filtering schemes. The snapshot approach has gained more acceptance than the other in recent times, so we concentrate on it here. It has the advantage of not having to make any questionable assumptions about how the system got to its present state. It matters only that the system is in a particular state "now," and the RAIM decision as to failure or no-failure is based on current observations only.

The theoretical structure for RAIM is statistical detection theory. Two hypothesis-testing questions are posed: 1) Does a failure exist? and 2) If so, which is the failed satellite? (It is usually assumed that there is only one failure at a time.) The answer to question 1 is sufficient for supplemental navigation because, presumably, there is an alternative navigation system to fall back on if a failure is detected. However, in the case of sole-means navigation, both questions 1 and 2 must be answered. Here, the errant satellite must be identified and eliminated from the navigation solution, so that the aircraft can proceed safely with an uncontaminated GPS solution. As might be expected, determining which satellite has failed is more difficult than simple failure detection, and it requires more measurement redundancy. Most of this section is devoted to the detection function as it pertains to the supplemental navigation application. Then there is a short discussion at the end of the section on failure isolation and its application in sole-means navigation.

The Radio Technical Commission for Aeronautics (RTCA) published its Minimum Operational Performance Standards (MOPS) for GPS as a supplemental navigation system in July 1991.[5] Table 1 shows its specifications for GPS integrity.

As indicated in Table 1, three distinct phases of flight are considered, and each has its own set of specifications. Failure here is defined to mean that the solution horizontal radial error is outside a specified limit, which is called "alarm limit."

Table 1 Radio Technical Commission for Aeronautics global positioning system integrity specifications for supplemental navigation

	Performance item			
Phase of flight	Alarm limit	Maximum allowable alarm rate	Time to alarm	Minimum detection probability
En route (oceanic, domestic, random, & J/V routes)	2.0 n.mi.	0.002/h	30 s	0.999
Terminal	1.0 n.mi.	0.002/h	10 s	0.999
RNAV approach, nonprecision	0.3 n.mi.	0.002/h	10 s	0.999

(Note that this is not exactly the same as saying that a particular satellite pseudorange error is outside a specified bound. In other words, failure is defined in the solution domain rather than the measurement domain.) The maximum allowable alarm rate in the table refers to the usual false alarm rate with no satellite malfunction. The 0.002 per hour figure is interpreted as meaning the false alarm probability associated with any independent sample will be no greater than 1/15,000. The detection probability is a conditional probability and is defined as (1 − miss probability). The specifications require that both the detection and alarm-rate specifications must be met globally at all times. Indirectly, this says that if, at any time or location, the satellite geometry is such that it cannot support both of these specifications simultaneously, then that geometry must be declared as inadmissible for integrity purposes. When combined, the two requirements are quite severe, and this causes the RAIM availability to be less than desirable for the nonprecision approach phase of flight.[6] Note that there is no firm availability requirement specified in Table 1.

Next, we look briefly at some of the major snapshot RAIM schemes that have been proposed over the past few years. The discussion in all cases is qualitative to the extent possible. Key references are cited, and the reader can refer to them for mathematical details.

II. Basic Snapshot Receiver Autonomous Integrity Monitoring Schemes and Equivalences

Three RAIM methods have received special attention in recent papers on GPS integrity. These are the following: 1) the range comparison method first introduced by Lee,[3] 2) the least-squares-residuals method suggested by Parkinson and Axelrad,[7] and 3) the parity method as described in the context of GPS in papers by Sturza and Brown.[8,9] All three methods are snapshot detection schemes because they assume that noisy redundant range-type measurements are available at a given sample point in time. Also, in all cases, the algebraic problem is linearized about some nominal value of vehicle position and clock bias. The basic measurement relationships are then described by an over-determined system of linear equations of the form

$$\boldsymbol{y} = \boldsymbol{G}\,\boldsymbol{x}_{\text{true}} + \boldsymbol{\epsilon} \tag{1}$$

where n = the number of redundant measurements; $\boldsymbol{y}$ = the difference between the actual measured range (or pseudorange) and the predicted range based on the nominal user position and the clock bias ($\boldsymbol{y}$ is an $n \times 1$ vector); $\boldsymbol{x}_{\text{true}}$ = three components of true position deviation from the nominal position plus the user clock bias deviation ($\boldsymbol{x}_{\text{true}}$ is a 4×1 vector); $\boldsymbol{\epsilon}$ = the measurement error vector caused by the usual receiver noise, vagaries in propagation, imprecise knowledge of satellite position and satellite clock error, selective availability, and, possibly, unexpected errors caused by a satellite malfunction ($\boldsymbol{\epsilon}$ is an $n \times 1$ vector); and $\boldsymbol{G}$ = the usual linear connection matrix arrived at by linearizing about the nominal user position and clock bias. ($\boldsymbol{G}$ is an $n \times 4$ matrix.)

This basic measurement equation is common to all the RAIM methods described presently. In doing so, it is convenient to use the six-in-view-case for tutorial purposes. The generalization to $n = 5$ or $n > 6$ is fairly obvious, so this is not discussed in detail.

A. Range Comparison Method

Imagine having six satellites in view. We would then have six equations in four unknowns. Now, suppose that we solve the first four equations (as if there were no noise) and obtain a solution that satisfies the first four equations. (The ordering of the six equations is immaterial.) The resulting solution could then be used to predict the remaining two measurements, and the predicted values could then be compared with the actual measured values. If the two differences (residuals) are small, we have near-consistency in the measurements, and the detection algorithm declares "no failure." On the other hand, if either or both of the residuals are large, it declares "failure." This is the essence of the range comparison method. It only remains to quantify what we mean by "small" and "large" and assess the decision rule's performance.

Proceeding further with the six-in-view example, we see that the two residuals that we would obtain with the range comparison method can be thought of as a two-tuple, and they represent a point in a test-statistic plane as shown in Fig. 1.

We now look for a decision rule that divides the plane into two distinct regions, one corresponding to the "no failure" hypothesis, and the other corresponding to the "failure" hypothesis. A common way to choose the decision boundary is to let it be an equal probability density contour, conditioned on the assumption of no satellite malfunction. If the statistics of the noise are Gaussian, the contour will be elliptical, as shown in Fig. 1, and the particular contour chosen is the one that sets the alarm rate at the desired value. For example, the alarm rate could be set at 1/15,000, as specified in the RTCA MOPS.[5] Setting the decision boundary quantitatively then firms up the decision rule, and all that remains is to assess the algorithm's performance under the alternative hypothesis. This is somewhat complicated, so we do not pursue it further here. The elliptical contour (or closed surface in hyperspace) makes the decision rule a bit awkward computationally. However, conceptually, the rule is simple: decide "no failure" if the test statistic lies inside the hypersurface; decide "failure" if it lies outside the hypersurface.

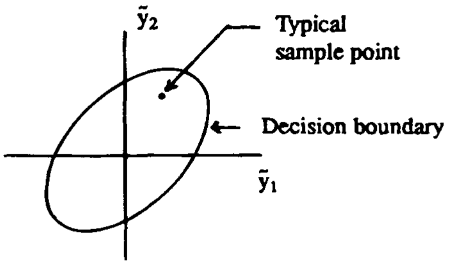


Fig. 1 Test statistic plane for the six-in-view case ($\tilde{y}_1$ and $\tilde{y}_2$ are the observed residuals).

B. Least-Squares-Residuals Method

Just as with the range comparison example, imagine a six-in-view situation where we have six equations in four unknowns. Now, instead of solving the first four equations as if there were no noise, say we look at the least squares "solution." This is well known and is given in Refs. 7 and 10.

$$\hat{x}_{LS} = (G^TG)^{-1}G^Ty \tag{2}$$

The least-squares solution can now be used to predict the six measurements in accordance with

$$(\text{predicted}\, y) = G\hat{x}_{LS} \tag{2a}$$

Six residuals are then formed in the measurement domain in much the same manner as was done in the range comparison method. The residuals can then be grouped together as a 6 × 1 vector, which we will call w. Substituting $\hat{x}_{LS}$ from Eq. (2) into Eq. (2a) then leads to the equation for w.

$$w = y - (\text{predicted}\, y) = [I - G(G^TG)^{-1}G^T]y \tag{3}$$

This is the linear transformation that takes the range measurement y into the resulting residual vector. (It can be easily verified that a similar equation also takes the measurement error ϵ into w.) In our example w is a six-tuple. However, Parkinson and Axelrad[7] show that there are constraints among the elements of w. For example, in the case at hand, if the elements of ϵ are independent zero-mean Gaussian random variables with the same variance, the sum of the squares of the elements of w has an unnormalized chi-square distribution with only two-DOF, rather than the six degrees that might be expected at first glance.

The sum of the squares of the residuals plays the role of the basic observable in the RAIM method under discussion, and Parkinson and Axelrad called it SSE (for sum of squared errors). We do also; i.e.,

$$\text{SSE} = w^Tw \tag{4}$$

This basic observable has three very special properties that are important in the least-squares-residuals decision rule.

1) SSE is a nonnegative scalar quantity. This makes for a simple decision rule. All we must do is partition the positive semi-infinite real line into two parts, one for "no failure," and the other for "failure." The dividing point is called the threshold.

2) If all elements of ϵ have the same independent zero-mean Gaussian distributions, then the statistical distribution of SSE is completely independent of the satellite geometry for any n. This makes it especially simple to implement a constant alarm-rate algorithm. All we must do is precalculate the thresholds (partitions) that yield the desired alarm rate for the various anticipated values of n. Then the real-time algorithm sets the threshold appropriately for the number of satellites in view at the moment.

3) For the zero-mean Gaussian assumption mentioned in 2, SSE has an unnormalized chi-square distribution with $(n - 4)$ degrees of freedom.

It should also be apparent that any other scalar variable monotonically related to SSE could also be used as the test statistic. Parkinson and Axelrad[7] suggested using $\sqrt{\text{SSE}/(n-4)}$ as the test statistic.[7] Later in Sec. III we also use $\sqrt{\text{SSE}/(n-4)}$ as our test statistic because this yields a linear relationship (rather than quadratic) between a satellite bias error and the associated induced test statistic. This is a convenience because the connection between the satellite bias error and the resultant radial position error is also linear.

To determine the threshold, we first use chi-square statistics and find the threshold for the value of n of interest, using SSE as the test statistic. This value can then be converted to the corresponding threshold for the $\sqrt{\text{SSE}/(n-4)}$ test statistic simply by dividing the SSE value by $(n-4)$ and square rooting the result. Such thresholds (rounded to integer values) are given in Table 2.

In summary, the least-squares-residuals RAIM method is especially simple in its implementation, because its test statistic is a scalar, regardless of the number of satellites in view. Calculating the test statistic involves some matrix manipulations, but these are no worse than calculating GDOP, PDOP, etc., which is done routinely as a background computation in most current GPS receivers.

C. Parity Method

The parity RAIM method as described by Sturza[9] is more formal and less heuristic than the other two methods. It is somewhat similar to Lee's range comparison method[3], except that the way in which the test statistic is formed is different. In the parity scheme, we first perform (conceptually) a linear transformation on the measurement vector $\boldsymbol{y}$ as follows:

$$\begin{bmatrix} \hat{\boldsymbol{x}}_{LS} \\ \text{---} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T \\ \text{---------} \\ \boldsymbol{P} \end{bmatrix} [\boldsymbol{y}] \tag{5}$$

Clearly, the upper partitioned part of the transformation yields the usual least-squares solution $\hat{\boldsymbol{x}}_{LS}$. The lower partitioned part, which yields $\boldsymbol{p}$, is the result of operating on $\boldsymbol{y}$ with a special (n $-$ 4) × n matrix $\boldsymbol{P}$ whose rows are mutually orthogonal, unity in magnitude, and also mutually orthogonal to the columns of $\boldsymbol{G}$. We ignore how $\boldsymbol{P}$ is found for the moment and concentrate on the $(n-4) \times 1$ vector $\boldsymbol{p}$, which is called the parity vector. (In our six-in-view example, $\boldsymbol{p}$ would be a two-tuple.) The very special way in which it is formed (i.e., as $\boldsymbol{Py}$) gives $\boldsymbol{p}$ some special properties. Under the usual assumption that the elements

Table 2 Approximate thresholds (test statistic = $\sqrt{\text{SSE}/(n-4)}$, noise σ = 33m, alarm rate = 1/15,000)

Number of satellites in view, n	Chi-square degrees of freedom	Threshold, m
5	1	132
6	2	102
7	3	90
8	4	82
9	5	77

of the random forcing function ϵ are independent similar zero-mean Gaussian random variables, the following statements can be made:

$$E[\boldsymbol{p}] = 0 \tag{6}$$

$$E[\boldsymbol{p}\boldsymbol{p}^T] = \operatorname{cov} \boldsymbol{p} = \sigma^2 \boldsymbol{I} \tag{7}$$

where σ^2 is the variance associated with any particular element of ϵ. Conceptually, in the parity method we use $\boldsymbol{p}$ as the test statistic. But, because of the special properties stated in Eqs. (6) and (7), we need not look at the individual elements of $\boldsymbol{p}$; they are all decoupled and have the same variance σ^2. For simple detection, we obtain all the information we need about $\boldsymbol{p}$ merely by looking at its magnitude, or its magnitude squared. Thus, in the parity method, the test statistic for detection reduces to a simple scalar, just as was the case with the least-squares-residuals method. Furthermore, Sturza shows that the sum of the squares of the elements of $\boldsymbol{p}$ and Parkinson and Axelrad's SSE[7] are identical,[9] i.e.,

$$\boldsymbol{p}^T\boldsymbol{p} = \boldsymbol{w}^T\boldsymbol{w} = \text{SSE} \tag{8}$$

This is to say that although the dimensionalities of $\boldsymbol{p}$ and $\boldsymbol{w}$ are different, their magnitudes are the same. The significance of Eq. (8) is simply this: if all we are interested in is the test statistic $\boldsymbol{p}^T\boldsymbol{p}$, then we do not actually have to go to the trouble of finding the orthogonal transformation $\boldsymbol{P}$ that leads to $\boldsymbol{p}$. Instead, we can just use SSE (ignoring $\boldsymbol{p}$ and $\boldsymbol{P}$) and get the same result as if we were to go to the work of finding $\boldsymbol{P}$, then forming $\boldsymbol{p}$ as $\boldsymbol{Py}$, and, finally, forming $\boldsymbol{p}^T\boldsymbol{p}$. Clearly, forming SSE directly in the measurement-residual space is easier.

In summary, in the detection application, the least-squares-residuals and parity methods lead to identical observables. Thus, with similar threshold settings, the two methods must yield identical results. Now, all that is needed to tie the range comparison method to the other two is to show that a linear transformation can be found that will always take the $(n - 4)$ vector range comparison test statistic into the parity vector $\boldsymbol{p}$. This is illustrated in Fig. 2.

It suffices here to say that such a transformation exists (see Ref. 11 for the derivation). Then, assuming that the decision boundaries in the two spaces are chosen to yield the same false alarm rate, it should be apparent that the range comparison decisions will be identical with those obtained from the parity and least-squares-residuals methods. The differences among the three detection meth-

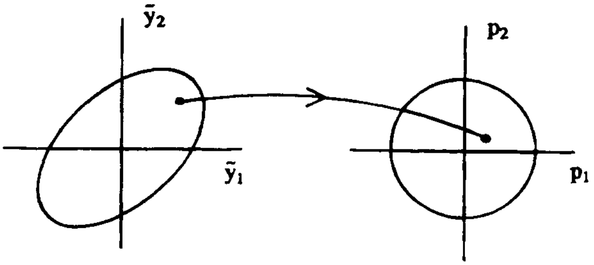


Fig. 2 Mapping from range-comparison space to parity space.

ods are, thus, primarily conceptual and a matter of computational convenience. For the sake of brevity, we simply refer to any of the three methods as the *parity detection method* henceforward.

D. Maximum Separation of Solutions

A RAIM method distinctly different from the parity method was suggested by Brown and McBurney in 1987.[12] It is probably the most heuristic of all the failure detection schemes. Assume that no more than one satellite has failed. Then, if there are n satellites in view, consider the n subset solutions obtained by omitting one satellite at a time from the full set. If a failure exists, the failed satellite will be omitted from one of the subsets, and the solution thus obtained will be a "good" solution. All other subset solutions will contain the failed satellite, and they will be in error to various degrees. Now, imagine the pseudorange error in the failed satellite gradually increasing with time. We would then expect the subset solutions to begin to spread apart with time, and a measure of this would be the maximum separation observed among the n solutions. (Note that one solution would remain near truth, because it does not contain the failed satellite.) On the other hand, if there is no failure present, the solutions should remain bunched around the true position. Thus, the maximum observed solution separation in the horizontal plane can be used as a test statistic. It is scalar and nonnegative, and it only remains now to set the threshold that separates the "no-failure" decision from the "failure" decision.

A. Brown[13] presented an interesting heuristic method of setting the threshold and assessing the radial error protected for the maximum separation scheme. It is illustrated for a five-in-view situation in Fig. 3. Suppose that under normal conditions we are assured that any four-satellite solution with a reasonable horizontal dilution of precision (HDOP) yields a solution within the 100-m 2-drms accuracy specification [with selective availability (SA)]. Then the maximum possible solution separation among the five solutions is 200 m, as shown in Fig. 3. Thus, let us set the test statistic threshold at 200 m. Then, any observed maximum separation greater than 200 m will be declared "failure." Furthermore, we are also assured that any horizontal radial error greater than 300 m will be detected because one of the solutions will always be within 100 m of truth, and

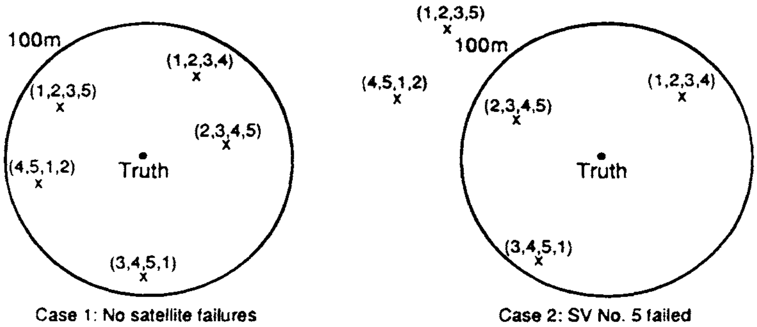


Fig. 3 Possible navigation solutions with five satellites.

the threshold on maximum separation has been set at 200 m. This establishes 300 m (and above) as the radial error being protected by the maximum separation scheme.

The preceding argument is approximate because 100-m 2-drms does not mean 100%. Also, the matter of false alarms has not even been considered. Nevertheless, the preceding reasoning led to radial-error-protection levels that were remarkably close to those obtained independently by Brown and McBurney using Monte Carlo simulation.[12,13] The results in both cases were somewhat optimistic, but this is understandable because the integrity specifications that were finally adopted by the RTCA group in 1991 were much more stringent than those that were being considered in 1987.

The maximum-solution-separation RAIM method is more difficult to analyze mathematically than is the parity method. This may account for the dearth of papers on the subject in the past few years. Much is now known about RAIM availability under conditions of the RTCA specifications and using the constant-false-alarm parity approach.[6] The same cannot be said for the maximum-solution-separation method.

E. Constant-Detection-Rate/Variable-Protection-Level Method

In September 1990, Brenner[14] presented a snapshot RAIM scheme that differs significantly from the four methods just presented. Brenner's approach starts out with the parity vector as the basic test statistic, and a threshold is set to yield the desired constant alarm rate, just as discussed previously. It is at this point that Brenner's method diverges from the others. In effect, he poses the question, "If we were to also keep the detection probability constant (at 0.999, for example) as the satellite geometry varies, what is the smallest radial error that we could protect and still stay within the desired specifications?" Brenner called this smallest-radial-error-protected the *detection level*, and it varies with the satellite geometry. (Detection level is also referred to as *protection radius* in some papers.) Furthermore, it can be calculated on-line (approximately, at least) on an essentially continuous basis in much the same manner as GDOP, PDOP, etc. This additional information is not required in the RTCA integrity specifications, but it is certainly permissible to display it in the cockpit as background information.

A typical plot of the calculated detection level as a function of time is shown in Fig. 4. The figure shows how the variable-detection-level method would be used within the context of the RTCA specifications where fixed, discrete alarm

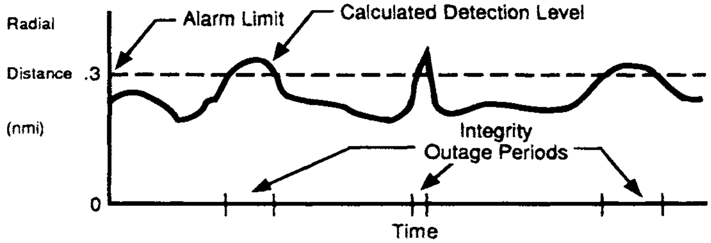


Fig. 4 Variable detection level (alarm limit for nonprecision approach is illustrated).

limits are set for each phase of flight. The calculated detection level is shown as the solid curve in Fig. 4. Whenever it exceeds the specified alarm limit (dashed line), the alarm would be annunciated. Assuming that the detection level is also displayed, it would obvious to the flight crew that the alarm was caused by inadequate satellite geometry and not a bona fide satellite failure. Of course, an alarm would also be annunciated if the test statistic exceeded its threshold, and this would be indicative of a satellite malfunction. The extent to which the variable-detection-level idea will be adopted by the aviation community remains to be seen.

III. Screening Out Poor Detection Geometries

The integrity requirements in the RTCA MOPS for GPS supplementary navigation are demanding.[5] They state that both the required detection probability and false alarm rate have to be met for *any* location and at *all* times. If the satellite geometry is such that both specifications cannot be met, then the alarm must be annunciated, indicating that integrity cannot be assured. Such geometries are referred to here as being *inadmissible.* It should be noted that these poor detection geometries might yield perfectly good navigation solutions; however, they simply do not have the appropriate redundancy to provide a good integrity check.

Inadmissible geometries detract from RAIM availability. Thus, it is important that they be screened out carefully. We do not want to throw out the good with the bad, so to speak. In early papers on GPS integrity, $\mathrm{HDOP}_{\max}$ (or $\mathrm{PDOP}_{\max}$) was frequently used as the screening criterion. The intuitive argument for this criterion is that with n satellites in view, consider the HDOPs associated with the n subset solutions obtained by omitting one satellite at a time. Let the largest of these HDOPs be denoted as $\mathrm{HDOP}_{\max}$. Now, if $\mathrm{HDOP}_{\max}$ is abnormally large, this means that the associated solution's projection onto the range of the missing satellite will be unreliable (i.e., noisy). This, in turn, makes it difficult to detect a failure (say, a modest bias error) on the missing satellite. A ceiling value is set on $\mathrm{HDOP}_{\max}$, and all geometries whose $\mathrm{HDOP}_{\max}$ exceed the ceiling value are declared inadmissible. One difficulty with this criterion is that it is not clear which of the DOPs should be used in the criterion. Perhaps GDOP would be more appropriate than HDOP, because the whole four-variable solution error reflects into projected range error on the missing satellite. Also, there is the matter of how a range error on a particular satellite projects into the horizontal solution, which is of primary interest in the avionics application. The net result of all this is that $\mathrm{HDOP}_{\max}$ is only a coarse measure of the quality of the detection geometry.

Two new (and better) methods of screening out poor detection geometries were introduced in the 1990–1991 period. One of these is called the $\delta H_{\max}$ method, and it comes from Sturza and A. Brown.[8,15] The other is referred to as the ARP method, and it was introduced independently by R. G. Brown et al.[16] It was shown later that the two methods are exactly equivalent for constant false-alarm-rate algorithms with similar thresholds.[11] These new methods of screening out poor detection geometries work out to be more exact than the older $\mathrm{HDOP}_{\max}$ criterion in terms of a sorting out the good geometries from the poor ones. This is demonstrated in Ref. 17.

The $\delta H_{\max}$ method is justified algebraically in Ref. 15, and it proceeds as follows.

1) Compute the HDOPs associated with the n subset solutions. Call these HDOP_i, $i = 1,2, \ldots, n$.

2) Compute the HDOP associated with the full n-satellite least-squares solution. Call it HDOP.

3) Then

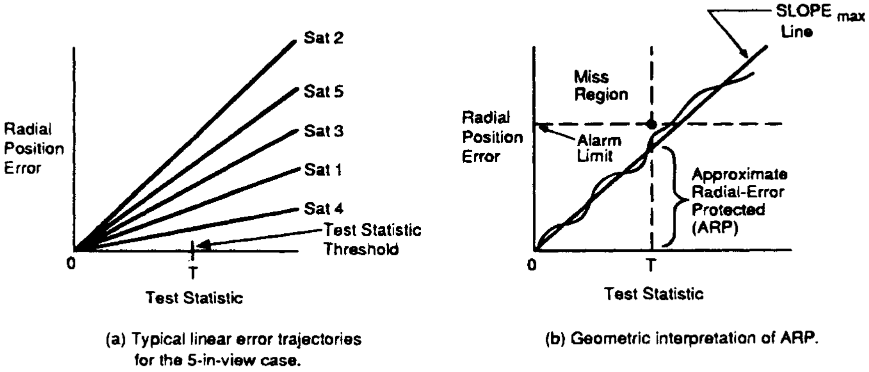
$$\delta H_{\max} = \text{Max}_i\,[\text{HDOP}_i^2 - \text{HDOP}^2]^{1/2} \tag{9}$$

The parameter $\delta H_{\max}$ then becomes an inverse measure of the quality of the satellite geometry for failure detection purposes (small values are best). A ceiling value for $\delta H_{\max}$ is then set in accordance with the integrity specifications, and if the calculated $\delta H_{\max}$ for the geometry at hand exceeds the ceiling value, the geometry is declared inadmissible; otherwise, it is admissible.

The ARP method is more intuitive than $\delta H_{\max}$ and it derives from geometric considerations when we look at a plot of position radial error vs test statistic. In the ARP method, it is convenient to work with a test statistic that is derived from SSE (or the magnitude of the parity vector squared) rather than SSE itself. Therefore, for screening purposes consider the test statistic to be

$$\sqrt{\text{SSE}/(n-4)} \tag{10}$$

With this definition, the radial position error and the test statistic have the same dimensions, and their ratio is dimensionless. It can be easily shown that a bias error on any particular satellite projects linearly into both the position-error and test-statistic domains. The slope, which relates the induced position error to the test statistic, can be readily calculated from the satellite geometry,[16] and it will be different for each satellite, as illustrated in Fig. 5a. For failure detection purposes, it should be clear that the satellite whose bias error causes the largest slope is the one that is the most difficult to detect. It is the one that produces the largest position error (which we want to protect) for a given test statistic (which is what we can observe). We call the slope associated with the most-difficult-to-detect satellite $\text{SLOPE}_{\max}$.



(a) Typical linear error trajectories for the 5-in-view case.

(b) Geometric interpretation of ARP.

Fig. 5 Radial error vs test statistic plots illustrating $\text{SLOPE}_{\max}$ and ARP.

Now suppose the threshold has been set to yield the allowed false alarm rate, and we then apply a ramp-type bias error to the $\mathrm{SLOPE}_{\max}$ satellite. If we had no noise to contend with, the radial error trajectory would move up along the linear $\mathrm{SLOPE}_{\max}$ line as shown in Fig. 5b. It is desirable, of course, that the trajectory not intrude into the miss region, or an undetected failure would occur. For the noiseless case just described, it can be seen that the alarm limit line could be moved down to the intersection of the $\mathrm{SLOPE}_{\max}$ line with the vertical threshold line. The ordinate of the intersection is the smallest radial error that we could protect under these ideal conditions (assuming that the threshold is fixed). We call this the approximate radial-error protected, or simply ARP. Clearly, from Fig. 5b it can be seen that ARP is given by

$$\mathrm{ARP} = \mathrm{SLOPE}_{\max} \times \text{threshold} \tag{11}$$

Now, to be more realistic, the noises on the other satellites cause the actual trajectory to follow a wavy line, which is also shown in Fig. 5b. The random deviation from the ideal linear trajectory will be large if the noises are large (e.g., with SA turned "on"), and vice versa. Therefore, as a practical matter, the alarm limit line must be kept comfortably above the ARP value to keep the wavy trajectory out of the miss region. For the current RTCA specifications, and with SA present, it has been found empirically that the actual radial error that can be protected is related to ARP by the following approximate equation.[11]

$$\text{Actual radial error protected} \approx 1.7 \times \mathrm{ARP} \tag{12}$$

This approximation is only within about 5%, but it is a useful rule of thumb to get a quick estimate of the position error that can be protected, given the calculated ARP figure for the satellite geometry at hand. (Note that the 1.7 multiplication factor would be considerably less with SA not present.)

Ceiling values for the ARP criterion have been determined by extensive simulation,[17] and they are repeated in Table 3. Note that these values are a function of the number in view as well as the phase of flight. Also, an exact relationship between ARP and $\delta H_{\max}$ is derived in Ref. 11, and it is

$$\mathrm{ARP} = \sqrt{n-4} \times \text{threshold} \times \delta H_{\max} \tag{13}$$

This equation may be used to convert the ARP ceiling values in Table 3 to corresponding $\delta H_{\max}$ ceiling values for those who prefer to work with $\delta H_{\max}$ rather than ARP. Note, however, that "threshold" in Eq. (13) is computed using

Table 3 ARP ceiling values

	Number of satellites in view		
Phase of flight	5	6	7 (or more)
Nonprecision approach	328	339	352
Terminal	1077	1135	1135[a]
Enroute	2159	2262	2262[a]

[a]These numbers are conservative estimated values. Very few seven-in-view geometries have ARP values this large.

$\sqrt{\mathrm{SSE}/(n-4)}$ as the test statistic rather than just SSE. Results given in Ref. 17 demonstrate that ARP (or equivalently, $\delta H_{\max}$) is more discriminating in screening out bad detection geometries than is $\mathrm{HDOP}_{\max}$. This is to say that the older $\mathrm{HDOP}_{\max}$ criterion unnecessarily eliminates some geometries that should, by rights, be admissible.

IV. Receiver-Autonomous Integrity Monitoring Availability for Airborne Supplemental Navigation

The results of an extensive study of RAIM availability for airborne supplemental navigation were presented by Van Dyke in 1992.[6] The ARP screening criterion was used in the study, and the ceiling values given in Table 3 (Sec. III) were used to distinguish between admissible and inadmissible geometries. The study involved five geographic areas: the continental United States (CONUS), the North Atlantic, Europe, the Central East Pacific, and the North Pacific, and a mask angle of 7.5 deg was used in the study. An abridged version of Van Dyke's results for unaided RAIM is presented in Table 4. The percentages in this table were obtained by averaging over the five geographic areas, and they have been rounded to the nearest 0.1%. Note that unaided RAIM availability for nonprecision approach is only about 94%, even for the best satellite configurations. This is obviously less than desirable.

The Van Dyke paper[6] also gives availability percentages for baro-aided RAIM over CONUS. The results are repeated here in Table 5. Fundamentally, any RAIM scheme is merely a consistency check among a group of redundant measurements. A baro-altitude measurement can be thought of as a range measurement to the center of the Earth, so it can also be brought into the group of measurements being used in the consistency check. Details about exactly how this is done are discussed in Sec V. For now, it suffices to compare the availability results for nonprecision approach with and without baro-aiding. Note the dramatic improvement obtained by adding baro-altitude information to the suite of measurements. For both 24-satellite constellations (i.e., 21 primary and optimized 21 + 3) the percent availability improves from 94% to 99%.

Table 4 Average unaided global positioning system receiver-autonomous integrity monitoring availability in percentages[a]

	Phase of flight		
Constellation	En route	Terminal	Nonprecision approach
21 Primary	99.7	99.4	94.4
21Primary–1 failure	99.1	98.2	87.5
21Primary–2 failures	97.7	95.7	82.0
21Primary–3 failures	96.3	93.2	73.9
Optimal 21	97.4	94.9	78.1
Optimized 21 + 3	99.8	99.5	93.8

[a] Ref. 6.

Table 5 Receiver-autonomous integrity monitoring availability over the continental United States with baro-aiding in percentages[a]

	Phase of flight		
Constellation	En route	Terminal	Nonprecision approach
21 Primary	99.99	99.45	99.02
21 Primary–1 failure	99.87	98.48	96.82
21 Primary–2 failures	99.79	96.42	93.90
21 Primary–3 failures	99.39	94.32	90.43
Optimal 21	99.96	95.34	96.13
Optimized 21 + 3	99.99	99.69	99.20

[a]Ref. 6.

V. Introduction to Aided Receiver-Autonomous Integrity Monitoring

It was mentioned in Sec. IV that non-GPS measurements can also be added to the suite of measurements being used for the RAIM consistency check. We concentrate here on snapshot range-type measurements such as might be obtained form GLONASS, baro-altitude, or Loran (with master stations synchronized to GPS time). Then other types of measurements such as inertial and time (clock coasting) are discussed briefly. We use GLONASS as an example of the use of non-GPS measurements in the RAIM consistency check. The methodology for incorporating baro-altitude or Loran measurements into the suite of measurements follows.

The fundamental measurement in GLONASS is pseudorange, just as in GPS. The main difference (as it affects RAIM) is the variance of the measurement error. Assume for the moment that our RAIM scheme uses one GLONASS measurement in addition to the usual suite of GPS measurements. This simply adds one linearized equation of the form

$$y_g = -C_{xg}\,\Delta x - C_{yg}\,\Delta y - C_{zg}\,\Delta z + \Delta T + \epsilon_g \tag{14}$$

where Δx, Δy, Δz, and ΔT are elements of the $\boldsymbol{x}$ vector; C_{xg}, C_{yg}, and C_{zg} are direction cosines between the user east, north, and up axis and the line of sight to the satellite; and ϵ_g is the measurement error of the non-GPS measurement. We assume here that the offset between GPS time and GLONASS time is known, so ΔT is only the receiver clock bias as before (in units of range). The y_g quantity of the left side of Eq. (14) is the usual linearized measurement: i.e., the difference between the actual measurement and the predicted value based on the nominal $\boldsymbol{x}$ about which the linearization takes place [see Eq. (1)]. For the sake of simplicity, let us assume that ϵ_g has exactly half the standard deviation of that associated with the GPS measurements. However, the previous unified RAIM theory discussed in Sec. II is based on the assumption of independent measurement errors, all having the same variance. Therefore, something has to be done to the GLONASS measurement equation to bring it into line with the other equations. It should be apparent that this can be accomplished by multiplying both sides of Eq. (14) by 2. The result is

$$2y_g = (-2C_{xg})\,\Delta x + (-2C_{yg})\,\Delta y + (-2C_{zg})\,\Delta z + (2)\,\Delta T + 2\epsilon_g \qquad (15)$$

where the coefficients that make up the extra row of the new G matrix are shown in parentheses for emphasis. The left side of Eq. (15) says that we must now consider the GLONASS linearized measurement scaled up by a factor of 2 as the "measurement" in our set of modified linear equations. In effect, this gives extra weight to the measurement residual on the GLONASS satellite. This is as it should be. According to our assumption, it is a more accurate measurement than the corresponding GPS measurements (with SA turned on).

This same measurement rescaling procedure can be followed for any other snapshot-type range or pseudorange measurement. It is worth mentioning that baro-altitude is a range measurement (in contrast to pseudorange), so the clock bias term does not appear in the measurement equation in this case. It should also be mentioned that this procedure does not apply (directly, at least) to nonsnapshot-type measurements; i.e., those that involve past as well as present measurement information.

Inertial navigation systems (INS) and GPS receiver clocks both have unstable error characteristics. This gives rise to some special problems when these sources of information are coupled into RAIM schemes. This can be illustrated for a receiver clock with the following scenario. Suppose that the aircraft has enjoyed a period of good satellite geometry during the first portion of the flight, and the clock bias and drift have been accurately calibrated during this period by the receiver Kalman filter. The RAIM integrity checks have indicated that everything is normal. Then comes a short period when the satellite geometry is not admissible for RAIM purposes. The question arises: "Why not merely treat the predicted clock bias (based on calibrations just prior to entering the bad geometry period) as a noisy measurement and add it to the measurement suite, thus making RAIM viable during the otherwise bad geometry period?" This would seem to make sense intuitively, provided, of course, that the clock is reasonably stable. The weakness in this reasoning is that the GPS measurements were used to calibrate the clock, and they could have drifted off considerably before triggering the alarm. Even for nonprecision approach, which is the most demanding phase of flight, the alarm limit is 0.3 n.mi. This is roughly an order of magnitude greater than normal GPS position error. The situation is even more exaggerated for the terminal and enroute phases of flight where the alarm limits are 1 and 2 n.mi, respectively. Thus, the clock calibration that we can actually rely on may not be as good as we might think initially. We are not assured that an errant satellite has not pulled the position and clock bias solutions considerably out of tolerance just prior to the bad geometry period. A similar problem would exist with INS-aided RAIM if the INS is continually updated with GPS which is normally the case. The basic problem is that the aiding sources are not independent of the satellite measurements that they are trying to verify. The beauty of aiding RAIM with baro-altitude, GLONASS, or Loran is that the errors in these sources are genuinely independent of the GPS satellite errors.

We will not try to predict how effective either INS or clock coasting might prove to be in the overall RAIM problem. The cautions just mentioned should be kept in mind, however. There have been a number of imaginative papers on the subject recently, and it is likely that there will be more as we get into the period when GPS will be used for sole-means navigation.[18,19]

VI. Failure Isolation and the Combined Problem of Failure Detection and Isolation

A. Introductory Remarks

All of the preceding discussion about RAIM has been directed toward the failure *detection* problem. This is sufficient for supplemental navigation where, presumably, there is an alternative navigation system to fall back on in case a failure is detected. Simple detection is not enough for sole-means navigation, however. There, the integrity system must also be able to isolate (i.e., identify) the errant satellite so that it can be removed from the navigation solution. This combined problem of failure detection and isolation (FDI) is often referred to simply as the FDI problem. The detection half of the problem has been studied in great detail over the past few years, and much is now known about RAIM performance for supplemental navigation.[6] Availability is the key measure of performance for supplemental navigation, because it is assumed that the detection and false alarm rate specifications will be met by properly screening out poor detection geometries.

The performance of RAIM for sole-means navigation has not been assessed as thoroughly as it has for supplemental navigation. There are two reasons for this. First, specific requirements for FDI have not been recommended by RTCA Special Committee 159 as yet. When these recommendations do arrive, it is likely that they will not be identical with those for supplemental navigation. For example, a false alarm is just a nuisance matter in supplemental navigation; on the other hand, in the sole-means case an unnecessary alarm that leads to a RAIM outage or an inferior navigation solution could be a serious safety matter.[20] Thus, the whole RAIM specifications matter must be reconsidered for sole-means navigation.

A second reason for lack of good performance data for FDI is that the methodology for solving the isolation half of the problem is still evolving. There is still plenty of opportunity for innovation on this problem, and new papers on FDI appear regularly at current navigation meetings. We do not attempt here to predict exactly where FDI research will lead ultimately. However, there are some basics of parity theory that will no doubt play a role in the final FDI solution, so we now continue the parity discussion that began in Sec. II.

B. Parity Method and Failure Detection and Isolation

Parity theory provides an especially useful geometric perspective in the FDI problem. Recall that we only use the magnitude of the parity vector (or a related quantity) as the test statistic for detection. We see presently that the direction of the parity vector provides useful information about the identity of the failed satellite.

Recall from Eq. (5) that the parity vector $\boldsymbol{p}$ is given by

$$\boldsymbol{p} = \boldsymbol{P}\boldsymbol{y} \tag{16}$$

where $\boldsymbol{y}$ is the measurement vector and $\boldsymbol{P}$ is the special transformation that takes us from the n dimensional measurement space to the $(n - 4)$ dimensional parity space. (We discuss the computation of $\boldsymbol{P}$ later in Sec. VI.C.) By definition, $\boldsymbol{P}$ has some very special properties.

1) The rows of $\boldsymbol{P}$ are orthogonal to the columns of $\boldsymbol{G}$.

2) The rows of $\boldsymbol{P}$ are mutually orthogonal (i.e., with each other).

3) The rows of $\boldsymbol{P}$ are normalized so that each of their magnitudes (i.e., Euclidean or 2-norm) is unity.

Note that if $\boldsymbol{y}$ in Eq. (16) is replaced with its equivalent $(\boldsymbol{G}\boldsymbol{x}_{\text{true}} + \boldsymbol{\epsilon})$, the orthogonality property dictates that $\boldsymbol{P}\boldsymbol{G}\boldsymbol{x}_{\text{true}} = 0$; and thus

$$\boldsymbol{p} = \boldsymbol{P}\boldsymbol{\epsilon} \tag{17}$$

In effect, the definition of $\boldsymbol{P}$ makes it such that the true value of $\boldsymbol{x}$ is blocked in projecting $\boldsymbol{y}$ into $\boldsymbol{p}$, and all we are left with is the measurement *error* projected into the parity space. Note that the dimension of the parity space is four less than that of the measurement space in this application. Thus, the projection of the measurement error into a useful test statistic is much easier to visualize in parity space than in the measurement-residual space (and with no loss of information).

Next, omit the measurement noise for the moment and consider the effect of a bias error b on a single satellite. For purposes of illustration, let us say that we have six-in-view, just as in the example in Sec. II. In this case, $\boldsymbol{P}$ is a 2 × 6 matrix. Suppose we put the bias on satellite four (i.e., the fourth element of the column vector ϵ). The resultant projection into parity space is then

$$\boldsymbol{p}(\text{for bias on sat. 4}) = \begin{bmatrix} p_{14} \\ p_{24} \end{bmatrix} b \tag{18}$$

(lower case $\boldsymbol{p}$ with *two* subscripts denotes elements of $\boldsymbol{P}$).

It can be seen from Eq. (18) that the parity vector induced by a bias on satellite four must lie along a line whose slope is p_{24}/p_{14}. This is shown in Fig. 6.

A similar argument applies to the parity vector that would be induced by placing a bias on any of the other five satellites. Each satellite has its own characteristic bias line, with a slope determined by the elements of the respective column vector of $\boldsymbol{P}$, i.e.,

$$\text{Slope of char. line for sat. } i = p_{2i}/p_{1i}\,, \; i = 1,2, \ldots, 6 \tag{19}$$

A decision rule for identifying the failed satellite is now obvious.

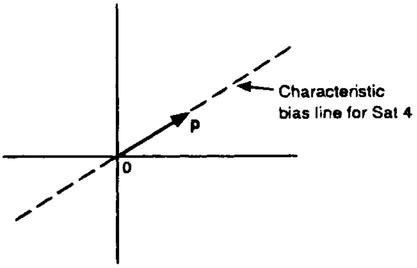


Fig. 6 Parity space showing the induced p caused by a bias on satellite 4 (no noise).

Decision rule: The failed satellite is the one whose characteristic bias line lies along the observed parity vector ***p***.

Let us now be more realistic and include noise in our FDI example. To illustrate the difficulty of meeting reasonably stringent (but hypothetical) specifications, suppose we say the alarm rate must be 1/15,000 (or less) and the desired detection probability is 0.999, just as in the supplemental navigation case. Furthermore, assume that we use a parity detection rule, and then follow it immediately with a "most-likely" isolation rule to identify the failed satellite. Most-likely here will mean the satellite whose characteristic bias line is closest to the observed parity vector. The probability of successful isolation of the failed satellite is to be at least as high as the detection probability of 0.999. For convenience, we will use $|\boldsymbol{p}|$ (or $\sqrt{\text{SSE}}$) as our test statistic.

Figure 7 shows the six characteristic bias lines for one of the six-in-view geometries used in the Van Dyke availability study.[6] The ARP value for this geometry is 1130, so it is admissible for detection purposes for both terminal and enroute flight (see Table 3). Now, suppose we consider a bias on satellite 4 that is just sufficient to trigger the alarm without noise. This induces a component vector in the parity domain that emanates from the origin and terminates on the alarm circle, as shown in Fig. 7. We then add to this a random Rayleigh noise vector (mainly caused by SA) whose mode is 33 m. The sum then terminates on the smaller dashed circle shown in Fig. 7. We know nothing about the direction of the random noise vector except that is uniformly distributed in angle. Now imagine performing a Monte Carlo experiment, keeping the bias and satellite geometry fixed and choosing the noise vectors at random. This would produce a cloud of data points centered around the center of the dashed (i.e., the noise)

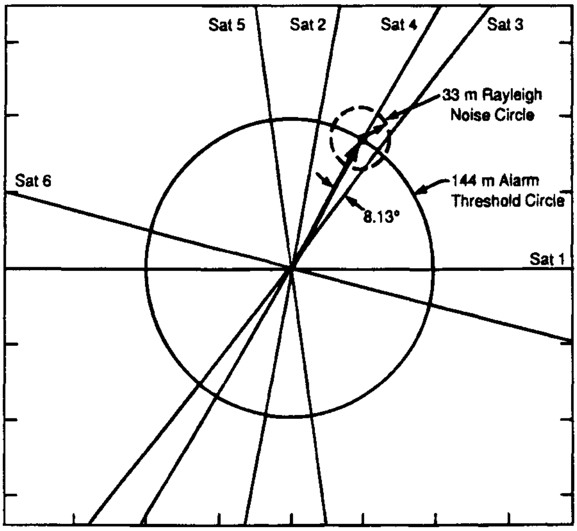


Fig. 7 Characteristic bias lines in parity space—six-in-view example (test stastic = $|\boldsymbol{p}|$, measurement noise σ = 33 m).

circle. Even without doing any calculations, we can see visually that a sizable fraction of the data would lie closer to the satellite 3 line than the satellite 4 line, and thus, these data would result in incorrect isolations. The fraction of misidentifications in this example would be about two orders-of-magnitude greater than desired! Thus, an FDI algorithm that simply tacks a single isolation appendage onto a detection stage does not even come close to satisfying reasonable specifications for this geometry.

As might be expected, there are certain equivalencies in the isolation part of FDI as well as in detection. Brown and Sturza[15] point out that the geometric "closest to the observed parity vector" decision rule can be replaced with an equivalent algebraic rule in the measurement-residual space. Which rule is to be used is purely a matter of computational convenience. The beauty of the parity approach lies in its vivid geometric interpretation. For example, it is obvious from Fig. 7 what must be done to improve the isolation performance. We must either 1) make the alarm threshold circle larger; or 2) make the noise circle smaller, or 3) add more satellites to the measurement suite to obtain a higher-order parity space (e.g., seven characteristic bias lines will diverge faster in three-space than six lines do in two-space). Also, the effect of the satellite bias error is clear from Fig. 7. Small biases are not a problem, because they (plus the noise) do not trigger the alarm. On the other hand, a very large bias will push the observed parity vector far from the origin, and a near-perfect identification will take place. Thus, very small and very large bias errors are not problems; the intermediate ones are those that are difficult to detect and isolate with certainty. All of these conclusions as to how various parameters interact in the FDI problem are obvious from one simple picture in parity space.

We do not pursue various avenues for improving the basic parity FDI scheme any further here. The main lesson to be learned from the preceding example is this. The direct snapshot use of the parity vector for both detection and isolation looks simple initially. Only after we put realistic numbers into the problem does the real difficulty becomes apparent. The tight specifications make this a truly difficult problem. It is likely that more sophisticated algorithms, such as the two-stage approach suggested in a recent paper by Lee,[20] will have to be developed before RAIM will become viable for sole-means navigation.

C. Calculation of the *P* Matrix

We now present two methods of computing the ***P*** matrix. One of these is formal and involves linear algebra theory; the other is less formal, but equally valid.

1. Formal Method

QR factorization decomposes a matrix into orthogonal and triangular factors. This is discussed in such works as Golub and Van Loan,[21] so we omit the details here. We begin with $\boldsymbol{G}$, the linear connection matrix that connects the state space to the measurement space [see Eq. (1)]. It is first factored in QR form; i.e.,

$$\boldsymbol{G} = \boldsymbol{QR} \tag{20}$$

where $\boldsymbol{Q}$ is the orthogonal factor. Consider next the transpose of $\boldsymbol{Q}$, which we denote as $\boldsymbol{Q}^T$.

Then

$$\boldsymbol{P} = \text{Bottom}(n - 4) \text{ rows of } Q^T \tag{21}$$

This completes the formal method of computing ***P***. (We note that *QR* factorization is one of the "built-in" functions in MATLAB®, which makes the formal approach especially easy for those who have access to MATLAB.[22])

2. *Heuristic Method*

A suitable ***P*** matrix that satisfies all the requirements stated in Sec. VI.B can be found by simply using these requirements as constraints in the choice of the elements of ***P***. (It is seen presently that one of these elements is arbitrary in our six-in-view example.) In words, an algorithm for finding ***P*** is as follows.

A. Begin with the bottom row of ***P*** (six elements):

1) Let element one be zero and element two be unity.

2) Use the orthogonal relationship between the rows of ***P*** and the columns of ***G***, and write four linear equations relating the remaining unknown elements of the last row of ***P*** to the respective elements of ***G***.

3) Solve the equations and write an unnormalized bottom row of ***P*** as $[0\ 1\ p'_{23}\ p'_{24}\ p'_{25}\ p'_{26}]$.

4) Now normalize the unnormalized row by multiplying by a scale factor that makes its magnitude unity. This yields the final bottom row of ***P***, which we denote with the usual notation $[0\ p_{22}\ p_{23}\ p_{24}\ p_{25}\ p_{26}]$.

B. Now go the next to the bottom row of ***P***.

1) Let element one be unity.

2) Use the orthogonality relationships between this row and ***G*** and the previously determined bottom row of ***P***, and write five linear equations in the remaining unknown elements of this row.

3) Solve the equations and write an unnormalized equation for the next to bottom row as $[1\ p'_{12}\ p'_{13}\ p'_{14}\ p'_{14}\ p'_{15}\ p'_{16}]$.

4) Now normalize the unnormalized row. This yields $[p_{11}\ p_{12}\ p_{13}\ p_{14}\ p_{15}\ p_{16}]$.

The determination of ***P*** is now complete. Note that there are 12 degrees of freedom in ***P*** and only 11 constraining equations. Thus, one element is arbitrary. By letting the first element in the bottom row be zero, we automatically place the characteristic bias line for satellite 1 on the zero-angle reference. Obviously, we could have put the zero in any of the other five positions, thus causing those respective bias lines to be on the reference axis. In doing so, we would not change the relative spacing among the bias lines. The effect would simply be a rotation of the whole characteristic bias line picture by an appropriate amount to bring the desired bias line to the zero-angle axis.

The extension of this algorithm to the seven-in-view and higher-order cases is straight-forward. For example, in the seven-in-view case, ***P*** is a 3 × 7 matrix with 21 elements. There are 15 orthogonal and 3 normalizing constraints that must be satisfied. Thus, 3 elements may set equal to zero (within some restrictions). A reasonable way to choose the zeros is to put them in left portion of the unnormalized $\boldsymbol{P}$ as follows:

$$\boldsymbol{P}(\text{unnormalized}) = \begin{bmatrix} 1 & p'_{12} & p_{13} & \dots & p'_{17} \\ 0 & 1 & p'_{23} & \dots & p'_{27} \\ 0 & 0 & 1 & \dots & p'_{37} \end{bmatrix} \tag{22}$$

This choice will force the first column vector of $\boldsymbol{P}$ to be aligned with the reference axis in 3-space; the second column vector will be normal to the third axis in 3-space; and the remaining columns will generally be vectors with three nontrivial elements. The extension to the higher-order cases is fairly obvious, so this is not pursued further.

One final comment is in order. The $\boldsymbol{P}$ matrix is not unique. However, the relative spacing among the characteristic bias lines is unique, and this does not change when a rotation or mirror-image reflection-type transformation is made that takes one $\boldsymbol{P}$ matrix into another equally valid $\boldsymbol{P}$ matrix.

D. Failure Detection and Exclusion Algorithm

Before leaving the subject of FDI, it should be mentioned that a variation on the basic FDI scheme just described has recently appeared in the literature.[23] This variation is usually referred to as FDE for failure detection and exclusion (in contrast to isolation). This is something of a play on words, but there are subtle differences that warrant treating FDE separately from FDI.

The basic idea of FDE can be illustrated with an eight-in-view example. Suppose that the normal detection suite operates with only six satellites (as suggested in Ref. 23), and an alarm occurs. Instead of immediately trying to isolate the bad satellite, the algorithm simply searches among all the other six-satellite subsets for one that will satisfy its self-consistency test, thus assuring that the bad satellite has been excluded. Two satellites are excluded, but the algorithm does not attempt to determine which of the two is the bad one. In this sense, the algorithm does not isolate the individual offender, it only separates it into a group of possible offenders. Furthermore, if the bad satellite just happened to be one of the pair that was excluded in the original suite of six prior to the failure, then there would be no alarm and no search would be necessary, an obvious fringe benefit. Contrast this with the usual FDI algorithm that would be using all eight satellites in its detection suite. It would be forced to detect and isolate the bad satellite and run the attendant risk of making a mistake.

Another attractive feature of the FDE philosophy is that it will accomodate multiple satellite failures, to a limited extent at least. For example, with eight-in-view and six satellites normally in the detection suite, two concurrent failures could be detected and excluded. The algorithm would have to search through all combinations of eight things taken six at a time (28) to do so, but this is routine. With nine-in-view, three failures could be detected, and so forth. All of this comes with a price, however. The computational effort escalates rapidly with the number in view, and it is still not clear how this extra capability affects the RAIM availability.

As mentioned earlier, the state of the art relative to both FDI and FDE is still evolving, and it has not been decided at this time exactly what RAIM scheme will be recommended by the RTCA committee studying the matter. However, parity space plays an important role in both FDI and FDE, so it is essential that those working with RAIM understand the fundamentals of parity-space methods.

References

[1]"1990 Federal Radio Navigation Plan," Document DOT-VNTSC-RSPA-90-3/DOD-4650.4, U.S. Depts. of Transportation and Defense.

[2]Kalafus, R. M., "Receiver Autonomous Integrity Monitoring of GPS," Project Memorandum DOT-TSC-FAA-FA-736-1, U.S. DOT Transportation Systems Center, Cambridge, MA, 1987.

[3]Lee, Y. C., "Analysis of Range and Position Comparison Methods as a Means to Provide GPS Integrity in the User Receiver," *Proceedings of the Annual Meeting of the Institute of Navigation* (Seattle, WA), June 24–26, 1986, pp. 1–4.

[4]Brown, R. G., and Hwang, P. Y. C., "GPS Failure Detection by Autonomous Means Within the Cockpit," *Proceedings of the Annual Meeting of The Institute of Navigation* (Seattle, WA), June 24–26, 1986, pp. 5–12.

[5]"Minimum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System (GPS)," Document RTCA/DO-208, Radio Technical Commission for Aeronautics, Washington, DC, July 1991.

[6]Van Dyke, K. L., "RAIM Availability for Supplemental GPS Navigation," *Proceedings of the 48th Annual Meeting of the Institute of Navigation*, June 29–July 1, 1992, Washington, DC.

[7]Parkinson, B. W., and Axelrad, P., "Autonomous GPS Integrity Monitoring Using the Pseudorange Residual," Navigation (Washington), Vol. 35, No. 2, 1988, pp. 255–274.

[8]Sturza, M. A., and Brown, A. K., "Comparison of Fixed and Variable Threshold RAIM Algorithms," *Proceedings of the Third International Technical Meeting of the Institute of Navigation*, Satellite Division, ION GPS-90 (Colorado Springs, CO), Sept. 19–21, 1990, pp. 437–443.

[9]Sturza, M. A., "Navigation System Integrity Monitoring Using Redundant Measurements," Navigation (Washington). Vol. 35, No. 4, 1988–1989, pp. 483–501.

[10]Brown, R. G., and Hwang, P. Y. C. "Introduction to Random Signals and Applied Kalman Filtering," 2nd ed., John Wiley & Sons New York, 1992.

[11]Brown, R. G., "A Baseline RAIM Scheme and a Note on the Equivalence of Three RAIM Methods," *Proceedings of the National Technical Meeting of the Institute of Navigation* (San Diego, CA), Jan. 27–29, 1992, pp. 127–137.

[12]Brown, R. G., and McBurney, P. W., "Self-Contained GPS Integrity Check Using Maximum Solution Separation as the Test Statistic," *Proceedings of the Satellite Division First Technical Meeting, The Institute of Navigation* (Colorado Springs, CO), 1987, pp. 263–268.

[13]Brown, A. K., "Receiver Autonomous Integrity Monitoring Using a 24-Satellite GPS Constellation," *Proceedings of the Satellite Division, First Technical Meeting, The Institute of Navigation* (Colorado Springs, CO), 1987, pp. 256–262.

[14]Brenner, M., "Implementation of a RAIM Monitor in a GPS Receiver and an Integrated GPS-IRS," *Proceedings of the Third International Technical Meeting of the Satellite*

Division of the Institute of Navigation, ION GPS-90 (Colorado Springs, CO), Sept. 19–21, 1990, pp. 397–406.

[15]Brown, A. K., and Sturza, M., "The Effect of Geometry on Integrity Monitoring Performance," *Proceedings of the Institute of Navigation Annual Meeting*, June 1990.

[16]Brown, R. G., Chin, G. Y., and Kraemer, J. H., "Update on GPS Integrity Requirements of the RTCA MOPS," *Proceedings of the 4th International Technical Meeting of the Satellite Division of the Institute of Navigation*, ION GPS-91 (Albuquerque, NM), Sept. 11–13, 1991.

[17]Chin, G. Y., Kraemer, J. H., and Brown, R. G., "GPS RAIM: Screening Out Bad Geometries Under Worst-Case Bias Conditions," *Proceedings of the 48th Annual Meeting of the Institute of Navigation*, June 29–July 1, 1992.

[18]Diesel, J. W., "A Synergistic Solution to the GPS Integrity Problem," *Proceedings of the National Technical Meeting of the Institute of Navigation* (Phoenix, AZ), Jan. 22–24, 1991, pp. 229–236.

[19]McBurney, P. W., and Brown, R. G., "Receiver Clock Stability: An Important Aid in the GPS Integrity Problem," *Proceedings of the National Technical Meeting of the Institute of Navigation* (Santa Barbara, CA), Jan. 26–29, 1988, pp. 237–244.

[20]Lee, Y. C., "Receiver Autonomous Integrity Monitoring (RAIM) Capability for Sole-Means GPS Navigation in Oceanic Phase of Flight," *IEEE 1992 Position Location and Navigation Symposium (PLANS) Record*, Monterey, CA, 1992, pp. 464–472.

[21]Golub, G. H., and Van Loan, C. F., *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.

[22]*386-MATLAB Users Guide*, The Math Works, South Natick, MA, Oct. 15, 1990.

[23]Van Graas, F., and Farrell, J. L., "Baseline Fault Detection and Exclusion Algorithm," *Proceedings of the 49th Annual Meeting of the Institute of Navigation* (Cambridge, MA), June 21–23, 1993.

Chapter 6

Integration of GPS and Loran-C

Per K. Enge*
Stanford University, Stanford, California 94305
and
F. van Graas†
Ohio University, Athens, Ohio 45701

I. Introduction

FROM 1945 through the 1970s, the Long Range Navigation System (Loran-C) was developed by the United States and the Soviet Union (the Soviet system is called Chayka) primarily for military use. In the 1970s, Loran-C was declared an official national system by the United States and Canada. These decisions spurred a wealth of civilian applications and tremendous commercial development. The integrated circuit and microprocessor improved the performance of Loran-C receivers, while greatly reducing their size and price. Solid-state Loran transmitters were introduced, and these had much greater reliability and efficiency than their vacuum tube predecessors.

Today, Loran provides service to nearly a million maritime, airborne, and terrestrial users throughout most of the northern hemisphere. The coverage of Loran is shown in Fig. 1, where the coverage contours are for the extremely reliable groundwave coverage. Skywave propagation does provide extended coverage, but it is not as reliable.

Recently, the United States has added four new Loran transmitters to cover the middle of the United States (to fill the "midcontinent gap").[1] With these new transmitters, five or more Loran signals are available 95% of the time over more than 95% of the conterminous United States (CONUS).

Furthermore, the Far East Loran Technical Group, which includes Japan, the Peoples Republic of China, Korea, and the Commonwealth of Independent States, has agreed to build and operate six Loran chains under the auspices of the International Association of Lighthouse Authorities. As part of this effort, Japan and Korea have already contracted to build three solid-state Loran transmitters to replace older tube-type transmitters.

*Professor (Research), Department of Aeronautics and Astronautics.

†Associate Professor, Department of Electrical and Computer Engineering.

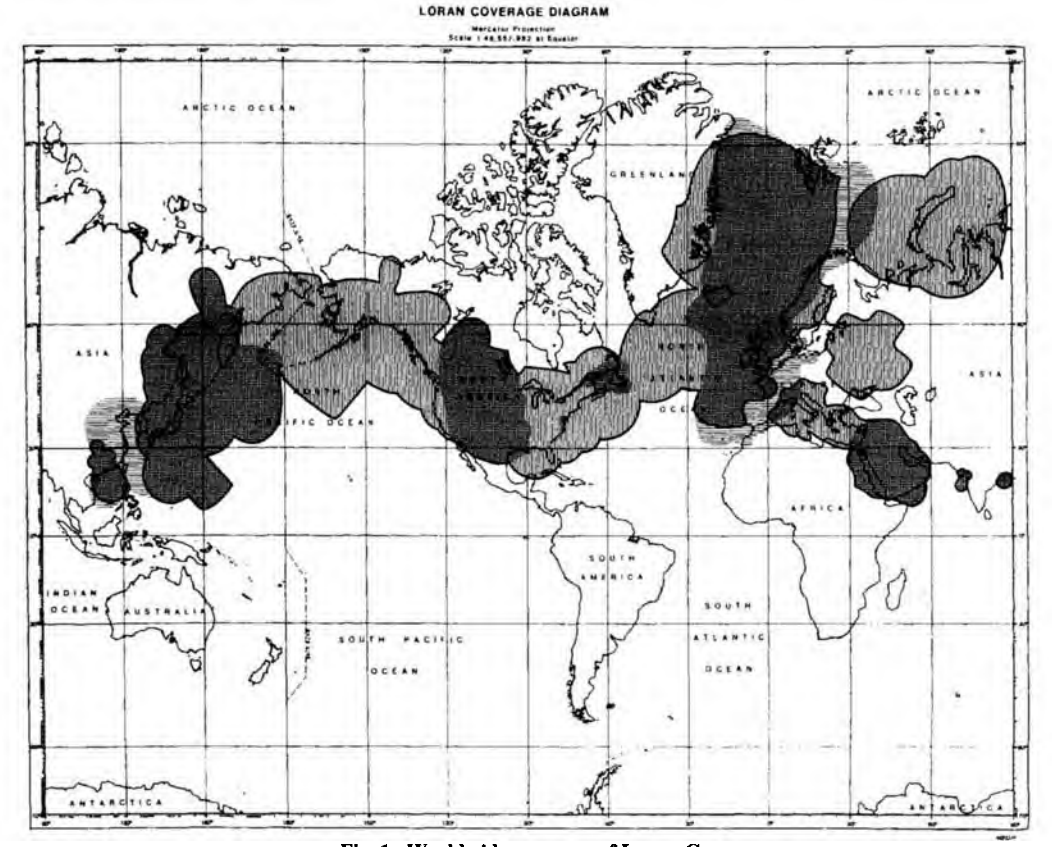


Fig. 1 Worldwide coverage of Loran-C.

In February of 1992, the Council of the European Communities adopted a decision on radionavigation systems for Europe, which stated "support efforts to set up a worldwide radionavigation system including European regional Loran-C chains with the purpose of enlarging worldwide Loran-C coverage in order to improve the safety of navigation and protection of the marine environment." As a consequence, nations in northwestern Europe are expanding their Loran coverage with modern transmitters.

In the summer of 1992, India installed two new Loran chains, which serve the waters near Bombay and Calcutta.

As described throughout this book, GPS is currently being installed and will be fully operational in the mid-1990s. These two vital radionavigation systems have much to offer each other.

A. Calibration of Loran Propagation Errors by GPS

Loran-C receivers measure the arrival time of groundwave radio signals emitted from a network of synchronized terrestrial transmitters. With repeatable accuracy measured in tens of meters, Loran provides excellent service to users navigating relative to landmarks that have been previously marked using Loran-C. However, its absolute (or geodetic) accuracy is currently limited to approximately one-quarter nautical mile (460 m) by uncertainties in the groundwave propagation speed. Fortunately, the Global Positioning System (GPS) can be used as a very accurate and convenient truth system to calibrate these uncertainties. Section III of this paper describes the results of such a calibration in the Gulf of Maine.

B. Cross-Chain Synchronization of Loran-C Using GPS

Loran-C transmitters are grouped into different chains, and each chain transmits groups of radio frequency pulses. Currently, Loran-C transmissions from different chains are not tightly synchronized. Consequently, Loran receivers can only compute time differences between signals from the same chain. If the receiver can receive many stations within a single chain, then this limitation is not troublesome. However, if the receiver is near the edge of coverage or if a transmitter has failed, then it can be helpful to remove this limitation and allow the receiver to compute "cross-chain" (or cross-rate) time differences. The time transfer capability of the GPS is one way of accomplishing the required cross-chain synchronization. In Sec. IV of this chapter, we examine (by example) the impact of cross-chain synchronization on Loran-C coverage.

C. Combining Pseudoranges from GPS and Loran-C for Air Navigation

Loran and GPS are currently widely used by aviators. Moreover, Loran-C is approved as a primary aid to navigation for certain airport approaches. However, Loran by itself does not qualify as a primary navigation system for terminal or enroute air navigation because one or more Loran signals may become unavailable. Signal loss may be attributable to "cycle slip," transmitter outage, or high noise. Furthermore, GPS by itself does not qualify as a primary civilian system because the time availability and integrity of stand-alone GPS are inadequate.

Section V of this chapter describes how information from the two systems can be combined in a hybrid receiver to give a system with improved coverage, availability and reliability.

Section II of this chapter provides an overview of Loran-C. Then Secs. III, IV, and V develop the above described techniques for using GPS and Loran together. Finally, Sec. VI is a brief summary.

II. Loran Overview

The Loran signal is shown in Fig. 2. As shown in Fig. 2a, Loran transmitters (also known as a Loran station or LORSTA) broadcast radio frequency pulses. Each pulse has a duration of approximately 200 μs and a center frequency of 100 KHz. The receiver identifies and tracks the arrival time of the sixth zero crossing of the pulse. This is the essential and fundamental measurement of any Loran receiver.

As shown in Fig. 2b, each transmitter periodically emits a group of eight (or nine) pulses. Loran transmitters are grouped into chains, and every transmitter in a given chain sends its group of pulses every group repetition interval (GRI), where the GRI varies from 50 to 100 ms. In fact, chains are distinguished by their unique GRIs. Each chain contains one master and two to five secondary stations, where the secondary transmissions are synchronized to the master transmission. A typical chain contains a master station *M* and three secondaries *X*, *Y* and *Z*. Secondary *X* emits its pulse group NED_X μs after the master transmits, where *NED* stands for nominal emission delay. Secondaries *Y* and *Z* are also synchronized and transmit NED_Y and NED_Z μs after the master.

As shown in Fig. 2c, phase codes control the polarity of the transmitted pulses. These phase codes distinguish the master signal from the secondary signals, and they repeat every two groups. The design of these sequences is described in an excellent article.[2]

A Loran-C user receiver measures the time difference (TD) between the arrival of the pulse groups from the master station and the secondary stations. The transmitter locations and nominal emission delays are well known, and the propagation speed of the Loran pulse can be estimated accurately. Therefore, each measured time difference defines a hyperbolic line of position (LOP) for the user. The intersection of two such LOPs defines the user's position in two dimensions. Loran is incapable of providing accurate estimates of altitude, because of the geometry of land-based transmitters relative to any user close to the surface of the Earth.

The transmitted Loran-C signal has a groundwave component that travels along the surface of the Earth, and a skywave component that is reflected off the ionosphere. The skywave component is not suitable for accurate position fixing because variations in the height and density of the ionosphere make the travel time of the skywave difficult to predict. Fortunately, the travel time of the groundwave is stable and predictable, and the design of the Loran pulse allows the receiver to separate the groundwave from the skywave. The groundwave attenuates as it propagates over Earth with finite ground conductivity. This attenuation limits each chain's range to approximately 500 nautical miles of the master station over land and 800 nautical miles over sea.

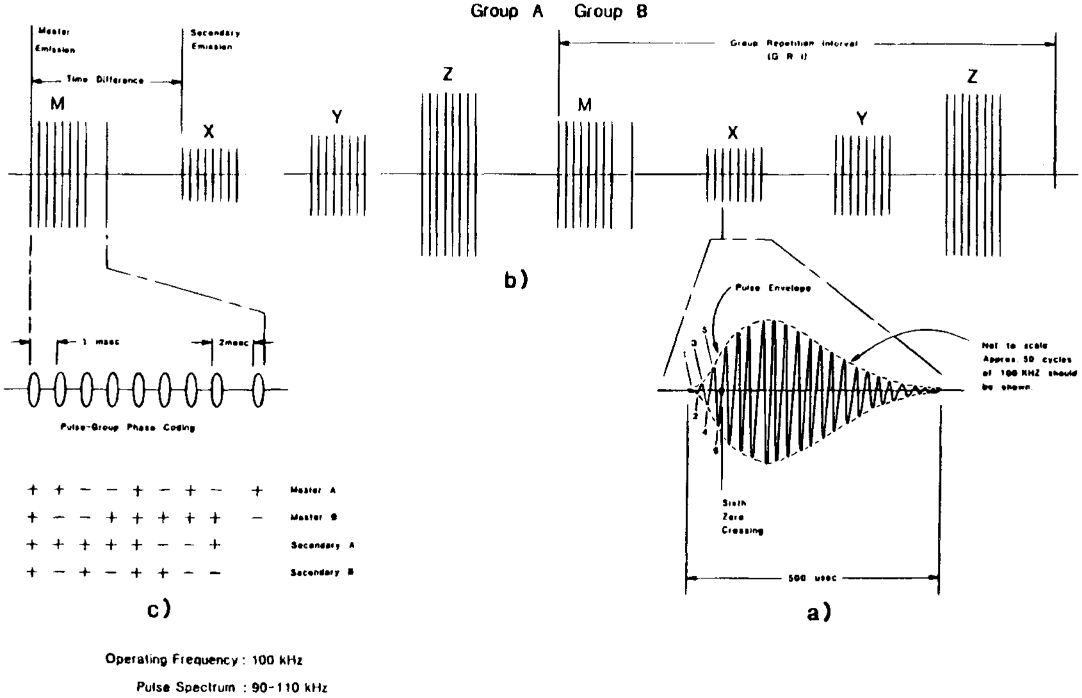

Group A
Group B
Master Emission
Secondary Emission
Time Difference
M
X
Y
Z
Group Repetition Interval (G R I)
M
X
Y
Z
b)
1 msec
2msec
Pulse-Group Phase Coding
+ + − − + − + − + Master A
+ − − + + + + + − Master B
+ + + + + − − + Secondary A
+ − + − + + − − Secondary B
c)
Pulse Envelope
1
2
3
4
5
6
Not to scale Approx. 50 cycles of 100 KHZ should be shown.
Sixth Zero Crossing
500 usec
a)
Operating Frequency : 100 kHz
Pulse Spectrum : 90-110 kHz

Fig. 2 Loran signal design: a) an individual Loran pulse, where the Loran reciever tracks on the sixth zero crossing of the 100 KHz carrier; b) illustration of how the individual pulses are grouped together with nine pulses for the master stations *M* and eight pulses for the secondary stations X, Y and Z; and c) illustration of the phase code that shifts the polarity of the pulses within a group. As shown, the master and secondary phase codes are unique. Additionally, the phase codes repeat every two groups.

Loran accuracy is limited by random errors and bias errors. Random errors are caused by noise and interference. They cause the time of the sixth zero crossing to vary rapidly relative to the time constant of the tracking loops in the receiver. Bias errors can also be caused by interference, but in the main they are caused by propagation effects. They cause the time of the sixth zero crossing to be offset and are difficult to remove by averaging. In other words, the decorrelation time of the propagation errors are long compared to the user platform dynamics.

Atmospheric noise is caused by lightning and is the most powerful natural noise source in the Loran band. The errors attributable to atmospheric noise are minimized by placing nonlinear signal processing elements in the receiver. These nonlinearities clip or limit the impulsive bursts characteristic of lightning noise.

Man-made signals near the Loran band are also important, particularly in Europe. The impact of this interference is minimized by careful design of the filters in the front end of the receiver. Interference at known frequencies can also be mitigated by careful selection of the group repetition intervals. Additionally, man-made interference and atmospheric noise are both reduced by making the time constants of the tracking loops as large as possible commensurate with the dynamics of the receiver platform. All told, atmospheric and man-made noise limit the short-term repeatable accuracy of Loran to tens of meters.

Propagation effects introduce errors that vary slowly or are bias errors. Indeed, the total travel time of the Loran-C signal (neglecting the difference between the transmitter and receiver clocks) is modeled as follows:

$$t_{\text{tot}} = t_{\text{PF}} + t_{\text{SF}} + t_{\text{ASF}} = t_{\text{SALT}} + t_{\text{ASF}} \tag{1}$$

The first term t_{PF} is known as the primary factor and is the travel time of a Loran-C signal moving at the speed of light in air with no boundary effects. The secondary factor t_{SF} is the additional time needed to travel over an all-seawater path. The "salt water model" $t_{\text{SALT}} = t_{\text{PF}} + t_{\text{SF}}$ is the total travel time over an all-seawater path. This term can be accurately modeled, because seawater variations have little effect on the travel time of the Loran pulse.[3]

The additional secondary factor t_{ASF} is the additional time needed to traverse any land segments. Moreover, t_{ASF} depends on the conductivity of the underlying ground and the overland distance traveled by the Loran pulse.[4–6] It grows to three or four μs after the Loran signal has traveled 200–300 km over typical land. The U.S. Defense Mapping Agency (DMA) publishes lookup tables for t_{ASF} for all U.S. Loran-C chains.[7] These tables give oversea estimates of t_{ASF} and are based on a model of ground conductivity in the US. With these data, the absolute accuracy (the accuracy of a position estimate with respect to the Earth's coordinates) of Loran is approximately 400 m.[8]

The additional secondary factor also varies with time as weather and climate change the effective conductivity of the land. Fortunately, these variations are small compared to the overall uncertainty in t_{ASF}. Consequently, Loran-C's repeatable accuracy (the accuracy with which a user can return to a position whose coordinates have been previously measured with Loran-C) is 18–90 m.[8]

III. Calibration of Loran Propagation Errors by GPS

Global Positioning System position fixes can be used to estimate t_{ASF} accurately, and thus to improve the absolute accuracy of Loran-C. In fact, Figs. 3 and 4

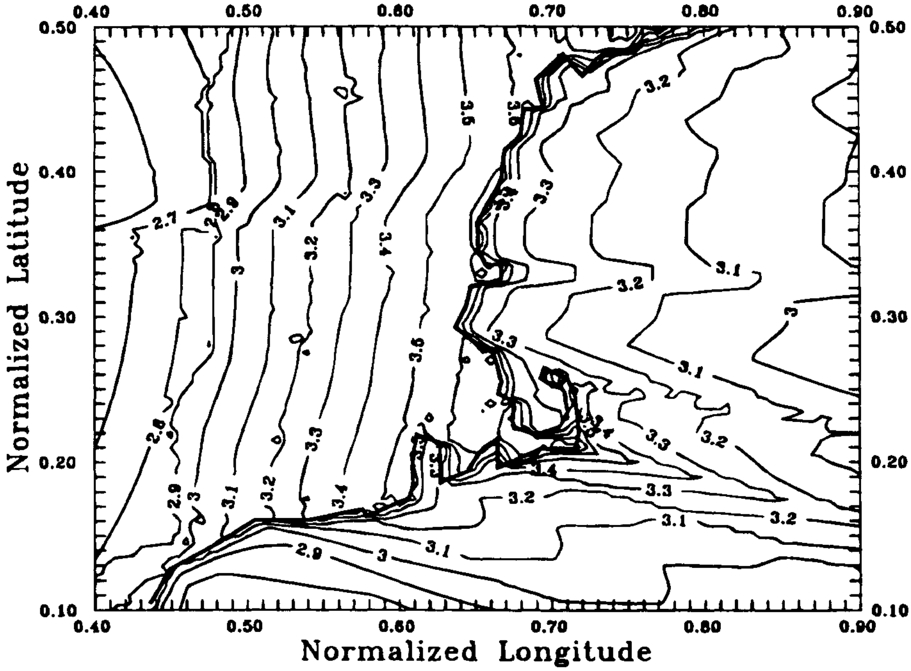


Fig. 3 Prediction of additional secondary phase factor (t_{ASF} in μs) for the Seneca–Nantucket time difference based on simultaneous GPS/Loran observations in the Gulf of Maine.

show t_{ASF} estimates in μs for New England based on GPS/Loran position fixes in the Gulf of Mexico. These figures are based on data collected by the U.S. Geological Survey (USGS) and processed as described by Refs. 9 and 10.

Briefly, the USGS collected Loran and GPS data in the Gulf of Maine over a 30-day period in October of 1985.[11] The USGS data include two-dimensional GPS position estimates and estimates of the GPS positional dilution of precision (PDOP). It also includes the Loran-C chain "9960" *M–X* and *M–W* time differences. The 9960 chain is identified by its group repetition interval—99.60 ms. The 9960 master station *M* is located in Seneca, New York, and secondaries *X* and *W* are located in Nantucket, Massachusetts and Caribou, Maine, respectively. The GPS measurements are used to compute t_{SALT} for each time difference in real time. Moreover, t_{SALT} is subtracted from the observed Loran travel time to produce an accurate estimate of $t_{\mathrm{ASF}} + \Delta NED$, where ΔNED is the error in the published nominal emission delay.

The algorithm developed by Refs. 9 and 10 uses the GPS-based observations of $t_{\mathrm{ASF}} + \Delta NED$ to calibrate a model that predicts t_{ASF} as a function of the conductivities for the *N* ground segments over which the signal propagates. The *N* conductivities and ΔNED can be adjusted to minimize the squared error between the model's output and the USGS observations.

Without calibration, the maximum and rms absolute errors of Loran in the Gulf of Maine are around 700 and 500 m, respectively, depending on the choice

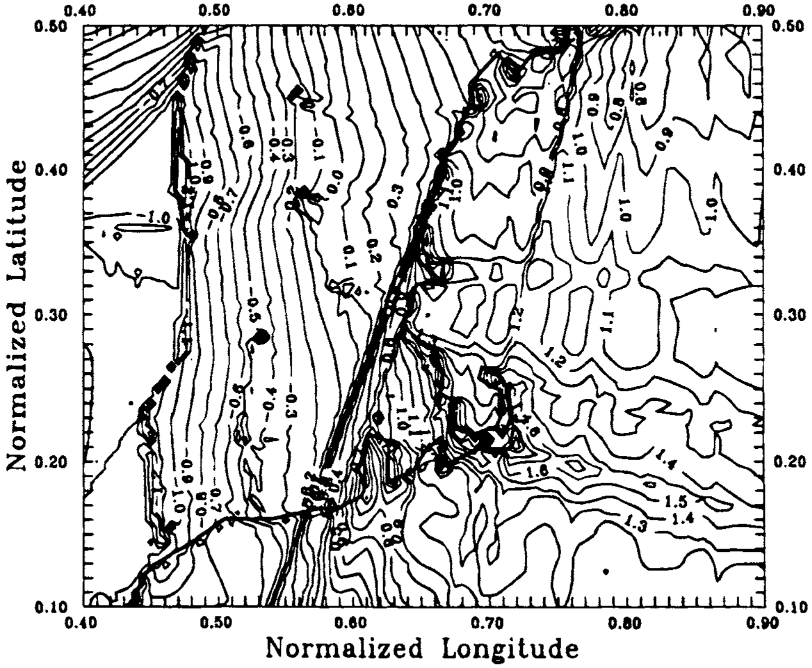


Fig. 4 **Prediction of additional secondary phase factor (t_{ASF} in μs) for the Seneca–Caribou time difference based on simultaneous GPS/Loran observations in the Gulf of Maine.**

of land conductivity. Significant improvements in the absolute accuracy of Loran can be achieved even with very simple calibrations. If the land conductivities are fixed a-priori and ΔNED (a single parameter) is optimized the maximum, and rms absolute errors fall to around 250 and 60 m, respectively. Alternatively, land can be treated as a single conductivity and this conductivity can be adjusted to reduce offshore additional secondary phase factor (ASF) errors. The performance of this practice results in maximum and rms errors of around 300–100 ms, respectively. More complicated approaches, which adjust multiple conductivities and ΔNED are also discussed in Refs. 9 and 10.

The GPS calibration can be used to form fixed correction tables (databases), which can be used to significantly improve the absolute accuracy of Loran for Loran-only missions. Indeed, a similar approach has been used to estimate t_{ASF} for the northern coast of Scotland.[12] Alternatively, a hybrid GPS/Loran receiver could perform such calibrations in real time. When GPS coverage is strong and integrity guaranteed, GPS could continously estimate t_{ASF}. The absolute accuracy of Loran calibrated in this way would be nearly equal to GPS and could be used to help an aircraft coast through a GPS outage.

IV. Cross-Rate Synchronization of Loran

Currently, the transmission times of the Loran masters are allowed to drift as long as they stay within ±2.5 μs of universal coordinated time (UTC). The transmission times of the secondary stations are more tightly controlled. They are controlled so that constant master–secondary time differences are maintained at certain system area monitors (SAM). For example, the Nantucket secondary station is controlled so that the Seneca–Nantucket time difference at Sandy Hook, New York is 26999.78 μs. This SAM control results in excellent repeatable accuracy for Loran users near the SAM.

In this section, we investigate the possible value of synchronizing the transmission from master stations in different chains. This analysis of cross-chain synchronization requires consideration of the following modes of receiver operation:

Single rate: The receiver can only use stations from a single GRI.

Two pair fixing: The receiver is also capable of using four stations, where two are in one GRI, and the other two are in a different GRI (a two pair fix).

Chain independent: The receiver can use any three or four stations regardless of which GRI they are in.

Single-rate receivers and receivers capable of cross-pair fixing exist today, and no change in chain timing is required to support these receivers. Chain-independent operation alone requires cross-chain synchronization.

Clearly, coverage for chain independent operation will always be at least as large as coverage for the other two receivers. However, is the chain-independent coverage ever significantly larger than coverage provided by the less flexible receiver? In other words, does a receiver that makes use of cross-chain synchronization provide more coverage than a receiver available today? To answer this, we plotted Loran coverage for an area that includes the Northwest United States and Southwest Canada (from 35 to 55°N and from 135 to 110°W). These coverage plots assume that the new North Central Chain and South Central Chain, which fill the midcontinent gap are operational. Finally, they assume that the station at Boise City is also dual rated with the Great Lakes Chain.

Figures 5 and 6 show coverage for the three modes of operation: single rate, two pair fixing, and chain independent. Both are for those noise levels exceeded 10% of the time annually. However all transmitters are healthy in Fig. 5, whereas the transmitter at George (47°N, 118°W) has failed in Fig. 6.

Both figures are for master independent receivers. Some Loran receivers are master dependent, whereas others are master independent. The master-dependent receivers must receive the signals from a Loran master station. They may require the master signal to identify which secondary is which, because the phase codes from Loran secondaries are identical and carry no identification information.

The figures assume that a signal is not useable if the received signal-to-noise ratio is less than −10 dB. They also assume that the Loran groundwaves suffer greater attenuation than predicted by an inverse distance law, because they are traveling over poorly conducting terrain. In fact, the figures assume that the conductivity of the underlying ground is 0.003 S/m. Figures 5 and 6 assume the atmospheric noise field strength is 49 dBμ (decibels above 1 microvolt per meter) when measured in a 20 kHz bandwidth. These noise levels are exceeded about 10% of the time annually in this geographical area as predicted by Ref. 13.

Figure 5 assumes that all of the stations are healthy. As shown in Fig. 5, the chain-independent receiver increases coverage slightly relative to two pair fixing in the western portion of the coverage area. The improvement is not dramatic, but under normal circumstances we should not expect it to be. After all, the

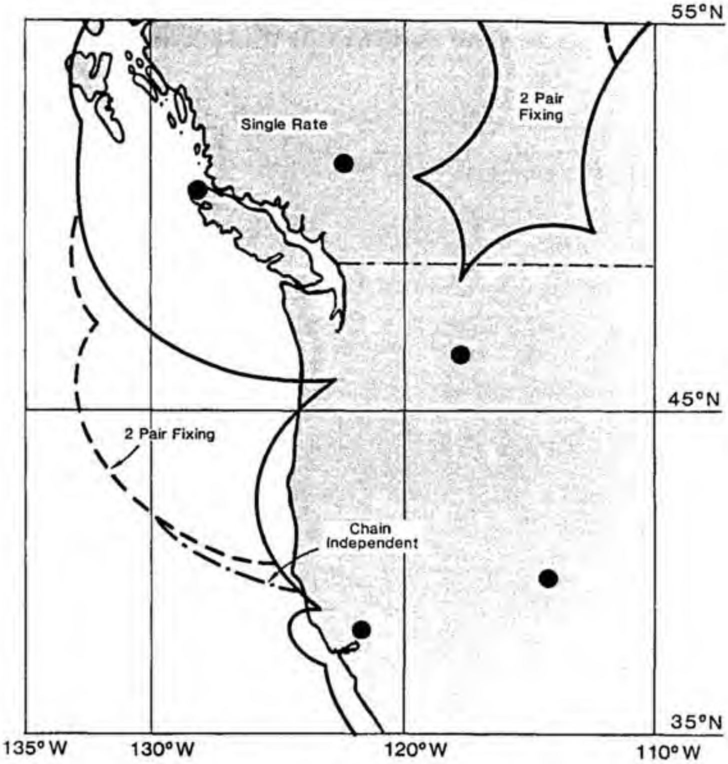


Fig. 5 Loran-C coverage in the U.S. Northwest for single-rate, two-pair fixing and chain-independent receivers. The noise level is exceeded 10% of the time annually, and all stations are healthy. The Loran transmitters are shown as dots.

Pacific, North Central and South Central chains have been designed to give excellent Loran coverage to all users in the western United States.

Figure 6 shows the coverage if the Loran station at George (Washington state) fails, and the noise field is 49 dBμ. Now, the chain-independent Loran receiver improves coverage dramatically. It seems that cross-rate synchronization would be very valuable in this case.

If a master station fails and the receiver is master dependent, then the advantage of cross-rate synchronization might not be as large. (The station at George is not a master). This follows, because the failure of a master would make all the stations in that rate unuseable. However, master-independent, single-rate receivers exist, and it seems likely that master-independent, chain-independent receivers can be realized.

As described in this section, cross-chain synchronization improves the availability of position fixing for a Loran-only receiver. However, combining GPS and Loran measurements in a combined receiver results in much more dramatic improvements. Such a receiver is described in the next section.

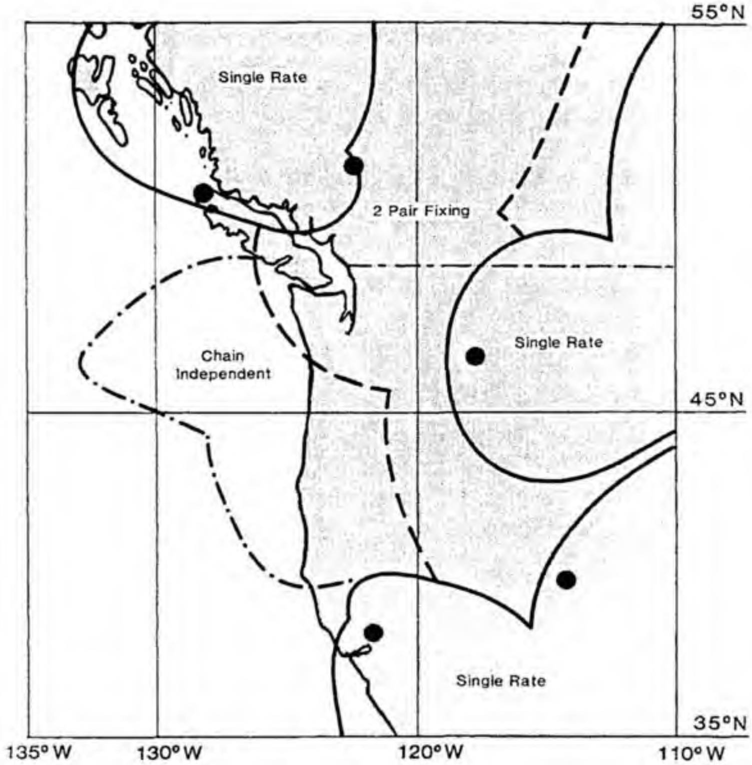


Fig. 6 Loran-C coverage in the U.S. Northwest for single-rate, two-pair fixing and chain-independent receivers. The noise level is exceeded 10% of the time annually. The Loran transmitters are shown as dots, but the station at George (47°N, 118°W) has failed.

V. Combining GPS Pseudoranges with Loran Time Differences

The planned availability of 98% with the GPS constellation is not sufficient for a primary navigation system. Furthermore, the availability of receiver autonomous integrity monitoring (RAIM) is significantly less than 98%. However, combining GPS with Loran-C in the user equipment significantly improves both the availability and the integrity of the position solution. Also, this type of integration provides dissimilar redundancy.

A. Navigation Equations

Global Positioning System and Loran-C measurement data can be combined in two ways: 1) GPS pseudoranges and Loran-C pseudoranges; and 2) GPS pseudoranges and Loran-C time differences (TDs). The first option potentially allows for the highest possible accuracy of the integrated solution.[14,15] To achieve this accuracy, it is necessary to synchronize the time of transmission of all Loran-C transmitters. In the United States, the time of transmission (TOT) of the

Loran-C master stations is held to within 100 ns with respect to UTC, but the transmissions of the secondary stations are under control of system area monitors. A SAM is located in the primary (marine) user area, and it adjusts the time of transmission of a secondary station so that the measured TD at the SAM is held to within ±50 ns of the controlling standard time difference (CSTD).[16] This provides a stable and accurate TD for users close to the line-of-position on which the SAM is located, but it causes the TDs at other locations to change as a function of varying propagation delays.[17] Furthermore, the Loran-C ASF corrections for TDs are available, whereas ASF corrections for pseudorange measurements do not exist at the present time. For these reasons, TDs should be used for Loran-C position calculations, unless the Loran chain uses TOT control.

A radionavigation range measurement is given by the following:

$$r_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} \tag{2}$$

where x, y, z is the three-dimensional user position; x_i, y_i, z_i is the position of the transmitting station; and r_i is the geometric range between the user and the station. A hyperbolic line of position is obtained by measuring the TD between the times of arrival of signals from two different transmitting stations:

$$\mathrm{td}_{i,j} = \left(\frac{b_{i,j} - r_i + r_j}{c}\right) + \mathrm{CD} \tag{3}$$

where $\mathrm{td}_{i,j}$ is the time difference observation for stations i and j; c is the speed of propagation of the radiowaves; $b_{i,j}$ is the geometric distance between the two stations; r_i and r_j are given by Eq. (2); and CD is the coding delay. The CD is constant, and is inserted by the secondary station to ensure that the transmissions of Loran stations do not overlap within the service area of the chain.

Note that for users at sea level, the Loran-C signals travel great-circle paths. To compensate for this, the transmitter locations are projected onto a locally level plane with respect to the user position estimate at distances equal to the great-circle distances to the transmitters.[18]

Next, the measurement equations are linearized to arrive at the position solution. An a priori estimate of the user position is used to form a Taylor series expansion, of which only the first-order terms are kept.

$$r_i = r_i + \left.\frac{\partial r_i}{\partial x}\right|_{x,y,z} \delta x + \left.\frac{\partial r_i}{\partial y}\right|_{x,y,z} \delta y + \left.\frac{\partial r_i}{\partial z}\right|_{x,y,z} \delta z \tag{4}$$

The a priori position estimate is used to calculate the estimate of the distance to the station r_i.

Equation (1) can now be linearized as follows:

$$\delta r_i = \begin{bmatrix} \frac{x - x_i}{r_i} & \frac{y - y_i}{r_i} & \frac{z - z_i}{r_i} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix} \tag{5}$$

If an unknown clock offset exists in the range measurement, the measurement is called a pseudorange:

$$pr_i = r_i + cb \tag{6}$$

where b is the unknown clock offset. Linearizing equation Eq. (5) results in a slightly different measurement equation:

$$\delta pr_i = \begin{bmatrix} \frac{x - x_i}{r_i} \frac{y - y_i}{r_i} \frac{z - z_i}{r_i} \; 1 \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \\ c\delta b \end{bmatrix} \tag{7}$$

A similar procedure is used to linearize the time difference equation:

$$\delta td_{i,j} = \underset{0}{\begin{bmatrix} \frac{x - x_j}{r_j} - \frac{x - x_i}{r_i} \\ \frac{y - y_j}{r_j} - \frac{y - y_i}{r_i} \\ \frac{z - z_j}{r_j} - \frac{z - z_i}{r_i} \end{bmatrix}}^{T} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \\ c\delta b \end{bmatrix} \tag{8}$$

Equations (6) and (7) relate a change in the user state to changes in the range and time difference measurements. In general, Eqs. (6) and (7) can be written as follows:

$$\delta y_i = \boldsymbol{g}_i \begin{bmatrix} \delta x \\ \delta y \\ \delta z \\ c\delta b \end{bmatrix} \tag{9}$$

where y_i is a measurement, and $\boldsymbol{g}_i$ is a row vector corresponding to that measurement. If all the measurements are included, Eq. (8) becomes the following:

$$\delta \underline{y} = G\delta \underline{x} \tag{10}$$

where $\underline{y}$ is a vector containing the measurements and $\underline{x}$ is the user state vector. G is a matrix containing data related to the geometry of the transmitting stations with respect to the user, as given by the row vectors $\boldsymbol{g}_i$.

Equation (9) can be used to solve for the user state vector iteratively based on the following steps:

1) Start with the user state estimate $\underline{\hat{x}}$ and the measurement vector $\underline{y}$.

2) Convert the Loran-C transmitter coordinates to a locally level plane with $\underline{\hat{x}}$ as the origin.

3) Convert the GPS satellite coordinates to the same locally level plane.

4) Calculate the estimated measurement vector $\underline{\hat{y}}$ using $\underline{\hat{x}}$, GPS satellite positions, and Loran-C transmitter positions.

5) Calculate the partial derivative matrix G; the rows of G are given in Eqs. (7) and (8).

6) Calculate the user state update from the following:

$$\Delta\underline{x} = (G^TG)^{-1}G^T(\underline{y} - \underline{\hat{y}}) \tag{11}$$

7) Update the user state as follows;

$$\underline{\hat{x}} = \underline{\hat{x}} + \Delta\underline{x} \tag{12}$$

8) If the magnitude of the update in step 7 is too large, go to step 4.

9) Use the new user state estimate in the locally level plane to update the user position in latitude, longitude, and height.

10) If the magnitude of the update in step 9 is too large, go to step 2.

11) Repeat steps 1–10 for the next set of measurements. (See Ref. 19 for a detailed description of this algorithm.)

To accomodate different measurement variances, Eq. (10) is left multiplied by a weighting matrix W[19]:

$$W\delta\underline{y} = WG\delta\underline{x} \tag{13}$$

In general, W could be derived from the measurement noise covariance matrix, but in most applications, it is sufficient simply to use a diagonal matrix, where the diagonal elements are the inverses of the measurement noise standard deviations.

B. Probability of Outage Results

This subsection quantizes the probability of outage (unity minus availability) for a hybrid GPS/Loran receiver. It presents a pair of figures that show the probability of outage for position fixing and autonomous fault detection.

For additional information, the reader is referred to Refs. 15 and 20. The first paper originally published the figures discussed in this subsection and gives a more detailed description of the underlying assumptions. It also includes results on the probability of autonomous fault isolation, which are not summarized here. The second paper also contains an excellent discussion of autonomous fault detection using Loran and GPS.

Probability of outage is the fraction of time/area for which the specified level of service is not available, and it is equal to 1 minus availability. For example, an availability of 0.999999 corresponds to a probability of outage of 10^{-6} or 0.0001%. This probability of outage means that service outages at an average location will last for approximately 30 s per year. An outage probability of 10^{-6} seems very low, but recall that hybrid GPS/Loran could serve as a primary air navigation system, and the requirements for such a system are very severe.

Figure 7 shows the probability of outage for position fixing vs the 2 drms accuracy of the all-in-view fix. If only GPS is used, then this probability is defined as follows.

$$Pr[2(\sigma_p)\text{HDOP} > \text{Accuracy}] \tag{14}$$

where HDOP is the horizontal dilution of precision of the GPS solution. If both systems are used, then the probability of outage is as follows

$$Pr[2(\sigma_p)\text{WHDOP} > \text{Accuracy}]$$

where WHDOP is the weighted HDOP for the combined solution.

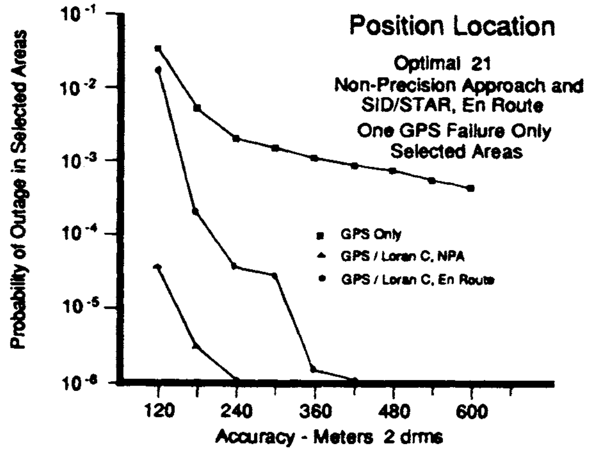


Fig. 7 Probability of outage for the position fixing service. This probability averages over the following underlying random variables: GPS satellite failures, LORSTA failures, atmospheric noise level, latitude, longitude, and time of day.

Both expression were evaluated with a GPS pseudorange error $1\sigma_p$ of 30 m, and this is a reasonable estimate for the (GPS) standard positioning service (SPS) with selective availability (SA). In both equations, probability is with respect to the following underlying random variables: GPS satellite failures, Loran Station (LORSTA) failures, atmospheric noise level, latitude, longitude, and time of day.

Figure 7 gives the probability of outage averaged over all of the selected geographical areas and time windows described in Ref. 15. It assumes that the receiver has full knowledge of all time offsets except its own clock bias. Additionally, Fig. 7 allows a single satellite shutdown from the optimal 21 constellation described in Ref. 21.

The top (worst) curve gives the outage probability for GPS position fixing, and the other two describe the performance of the GPS/Loran hybrid system. The best hybrid curve is for nonprecision approach and assumes that additional secondary phase factor (ASF) is well known, because of a Loran monitor at the destination airport. The middle curve is for the enroute case and assumes that ASF-related errors increase by 1ns/km (1σ) of range from the transmitters. These larger ASF error prevents the Loran aiding from providing as much improvement at high-accuracy levels.

Figure 8 gives probability of outage for the autonomous fault detection capability described in Chapter 5 of this volume. In contrast to position fixing, autonomous fault detection requires that all the signal subsets created by deleting a single signal have good geometry. Consequently, if GPS alone is used to detect GPS malfunctions (or faults), then we define the outage probability as follows:

$$Pr\{\max_{k\in S} \mathrm{HDOP}(k) > X\} \tag{15}$$

where HDOP (k) is the HDOP when satellite k is deleted, and S is the set of satellites in view. Hence, $\max_{k\in S}$ HDOP (k) is the maximum HDOP of the satellites that remain when each satellite in view is deleted one at a time.

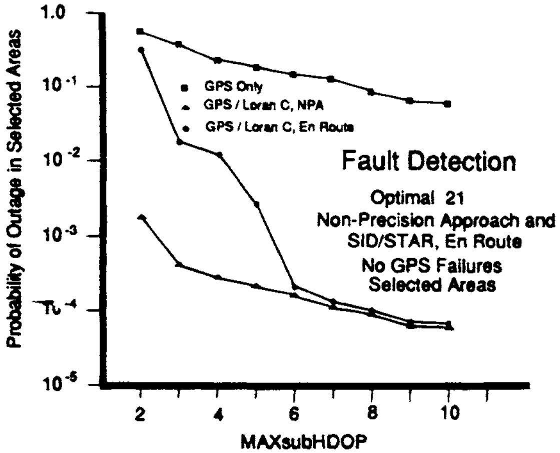


Fig. 8 Probability of outage for the autonomous fault detection fixing service. This probability averages over the following underlying random variables: GPS satellite failures, LORSTA failures, atmospheric noise level, latitude, longitude, and time of day.

If GPS and Loran are used to detect GPS malfunctions, then the probability that soft fault detection is unavailable is given by the following:

$$Pr\{\max_{k\in S} \text{WHDOP}(k) > X\} \tag{16}$$

where WHDOP (k) is the WHDOP when satellite k is deleted. In this case, the deleted transmitter only comes from the set of GPS satellites because Loran stations broadcast "aviation blink," which warns the user if any Loran signals are outside of specification.

Figure 8 shows the independent variable X from equations YY and ZZ on the horizontal axis. If the probability of false alarm and probability of missed detection are specified, then X can be related to the "protection limit" provided by the fault detection algorithm.[22-25] In general, Brown and Schmid[26] suggest that X is roughly equal to the desired protection limit in meters divided by 100. Figure 8 shows that Loran aiding does greatly reduce the probability of outage for fault detection. The improvement is greatest for a system that enjoys ASF calibration at the destination airport.

Although Figs. 7 and 8 assume TOT control for Loran-C, the results for TD operation should be very similar. The reason for this is that in the United States, almost all Loran-C transmitters are dual-rated, which means that one transmitter participates in two different chains.

VI. Summary

In this chapter, we have discussed and analyzed the substantive benefits of integrating GPS and Loran. First, GPS may be used to calibrate the propagation uncertainties that traditionally have limited the absolute accuracy of Loran-C.

Second, GPS time transfer can be used to synchronize transmissions from Loran transmitters in different chains. This cross-chain synchronization enables the measurement of cross-chain time differences by the Loran receiver, and these measurements could be very valuable if the Loran system is stressed by high noise or station failures.

Finally, GPS pseudoranges can be combined with Loran time differences in the user equipment. This combination greatly improves the availability of high-accuracy position fixing and autonomous fault detection and isolation. Of course, this improvement is especially dramatic in areas and times where and when GPS coverage is weak. Moreover, it is not very sensitive to whether or not there is an unknown time offset between GPS and Loran. The improvement in availability is greatest when errors attributable to Loran propagation uncertainties have been controlled through some means.

References

[1]Sedlock, A. J., "Mid-continent Loran-C Expansion," *Proceedings of the IEEE 1986 PLANS—Position, Location, and Navigation Symposium* (Las Vegas, NV), Institute of Electrical and Electronics Engineers, NY, Nov. 1986.

[2]Frank, R. L., "Current Developments in Loran-C," *Proceedings of the IEEE,* Vol. 71, Oct. 1983, pp. 1127–1139.

[3]McCullough, J., Irwin, B., and Bowles, R., "Loran-C Latitude–Longitude Conversion at Sea: Programming Considerations," *Proceedings of the Eleventh Annual Technical Symposium of the Wild Goose Association,* Washington, DC, 1982.

[4]Van der Pol, B., and Bremmer, H., "Diffraction of Electromagnetic Waves from an Electrical Point Source Round a Finitely Conducting Sphere." *Philosophical Magazine,* Vol. 7, No. 24, 1937, p. 825; No. 25, 1939, p. 817.

[5]Samaddar, S. N., "The Theory of Loran-C Ground Wave Propagation—A Review." *Navigation,* Vol. 26, No. 3, 1979, p. 173.

[6]Brunavs, P., "Phase Lags of 100 kHz Radiofrequency Ground Wave and Approximate Formulas for Computation," written communication, 1977.

[7]"Loran-C Correction Table for the Northeast U.S.A. (9960)," DMA Stock LCPUB2211200-C, 1988.

[8]"Federal Radionavigation Plan," DOT-TSC-RSPA-88-4, Dec. 1988.

[9]Pisano, J. J., "Using GPS to Calibrate Loran-C," Master's Thesis, Worcester Polytechnic Institute, Worcester, MA, Aug. 1990.

[10]Pisano, J. J., Enge, P. K., and Levin, P. L., "Using GPS to Calibrate Loran-C," *IEEE Transactions on Aerospace and Electronics Systems,* Vol. 27, No. 4, 1991.

[11]Irwin, B., and McCullough, J. "Gulf of Maine GPS/Loran-C Measurement Set," private communication, 1989.

[12]Last, D., and Ward, N., "The Use of DGPS for Mapping Loran-C Additional Secondary Phase Factors," First International Symposium on Real-Time Differential Applications of the Global Positioning System," Braunschweig, Germany, Sept. 1991.

[13]Spaulding, A. D., and Washburn, J. S., "Atmospheric Radio Noise: Worldwide Levels and Other Characteristics," National Telecommunications and Information Administration, NTIA Rept. 85–173.

[14]Van Graas, F., "Sole Means Navigation through Hybrid Loran-C and GPS," *Navigation,* Vol. 35, No. 2, 1988.

[15]Enge, P. K., Vicksell, F. B., Goddard, R. B., and Van Graas, F., "Combining Pseudoranges from GPS and Loran-C for Air Navigation," *Navigation,* Vol. 37, No 1, 1990.

[16]"Loran-C User Handbook," U.S. Department of Transportation, United States Coast Guard, Commandant Publication P16562.6, Washington, DC, 1992.

[17]Vicksell, F. B., and Goddard, R. B., "Implementation and Performance of the TOT Controlled French Loran Chain," *Proceedings of the Fifteenth Annual Technical Symposium of the Wild Goose Association,* New Orleans, LA, Oct. 21–24, 1986.

[18]Van Graas, F., "Hybrid GPS/Loran-C: A Next Generation of Sole Means Air Navigation," Ph.D. Dissertation, Nov. 1988. Athens, OH, Ohio University.

[19]Lawson, C. L., and Hanson, R. J., *Solving Least Squares Problems,* Prentice-Hall, Englewood Cliffs, NJ, 1974.

[20]Brown, R. G., and McBurney, P. W., "Loran-Aided GPS Integrity," *Proceedings of the Satellite Division First International Technical Meeting, Institute of Navigation* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 1988.

[21]Green, G. B., Massatt, P. D., and Rhodus, N. W., "The GPS 21 Primary Satellite Constellation," *Navigation,* Vol. 35, No. 1, 1988.

[22]Brown, A. K., "Receiver Autonomous Integrity Monitoring Using a 24-Satellite GPS Constellation," *Proceedings of the Satellite Division First Technical Meeting, Institute of Navigation* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 1987.

[23]Brown, R. G., and McBurney, P. W., "Self-Contained GPS Integrity Check Using Maximum Solution Separation as the Test Statistic," *Navigation,* Vol. 35, No. 1, 1988.

[24]Kalafus, R., and Chin, G. Y., "Performance Measures of Receiver Autonomous GPS Integrity Monitoring," *Proceedings of the 1988 National Technical Meeting of the Institute of Navigation* (Santa Barbara, CA), Institute of Navigation, Washington, DC, 1988.

[25]Lee, Y. C., "Performance Analysis of Self-Contained Methods for GPS Integrity Function," MITRE Corporation TR MTR-88W89, Nov. 1988.

[26]Brown, A. K., and Schmid, T., "Integrity Monitoring of GPS Using a Barometric Altimeter," *Proceedings of the National Technical Meeting of the Institute of Navigation* (Santa Barbara, CA), Institute of Navigation, Washington, DC, Jan. 1988.

Chapter 7

GPS and Inertial Integration

R. L. Greenspan*
Charles Stark Draper Laboratories, Cambridge, Massachusetts 02139

THE Global Positioning System (GPS) and inertial navigation systems (INS) have complementary operational characteristics. Even a modest attempt to combine their functionality in an integrated navigation system can produce a system performance superior to either one acting alone. However, because of the costs of such benefits, it is fitting to inquire about trade-offs that would justify the investment. Trade studies typically address the following questions:

1) What benefits of GPS/inertial integration are important in the application being considered?

2) What configuration of data paths (integration architecture) is appropriate for the application?

3) How complex are the integration algorithms required to provide the desired level of performance, with options for growth to meet future requirements?

This chapter devotes one section to address each one of these questions. Because of space limitations, the presentation is qualitative, with only limited recourse to the underlying mathematical structures required to understand integration filtering and the performance evaluation of an integrated navigation system. Wherever possible, the reader is directed to other chapters in this text for those details, or to the literature. Furthermore, the properties of GPS user equipment (UE) and inertial navigation systems cited here are generic rather than specific, and they are representative of technology circa 1993.

I. Benefits of GPS/Inertial Integration

The design of any complex navigation system for civilian or military markets reflects the designer's judgment of the best trade-off among the following factors:

1) Cost
   a) Development (nonrecurring)
   b) Life-cycle (recurring)
2) Installation constraints
   a) Volume, weight, power consumption

*Director, Electrical Design and Sensor Development Directorate.

b) Interfaces
3) Performance
a) Mission requirements/mission environment
b) Reliability/graceful degradation
c) Options for improvement

The following remarks emphasize the performance considerations because that is the area where the benefits of GPS/Inertial integration are most evident. However, cost and installation factors are often decisive. These are raised throughout the chapter wherever they are a significant differentiator between alternative integration techniques. Ultimately, the system designer must justify his or her design as being the best way to satisfy the design problem. It is of critical importance that the authorities who are managing the design team surface all requirements and constraints, both present and anticipated, so that informed and timely choices can be made among the alternatives.

The GPS system can provide a suitably equipped user with a position, velocity, and time (PVT) solution whose errors are generally smaller than those of any alternative navigation system. This performance is achieved in all weather, at any time of the day, and under specified conditions of radio-frequency interference, signal availability, and vehicle dynamics. Why then would we undertake the cost and complexity of integrating GPS UE with any other navigation sensor, and, in particular, with an inertial navigator?

The goal of integration is to provide more robust, and possibly more accurate, navigation service than is possible with stand-alone sensors. In particular, integration may be the only way to achieve the following[1-6]:

1) Maintain a specified level of navigation performance during outages of GPS satellite reception.

2) Provide a complete six-degree-of-freedom navigation solution (translational and rotational motion) at a higher output rate than is conventionally available from GPS alone.

3) Reduce the random component of errors in the GPS navigation solution.

4) Maintain the availability of a GPS solution in the presence of severe vehicle dynamics and interference.

Most civilian and non-Department of Defense Government GPS users have access to the Standard Positioning Service (SPS) only, which is subject to intentional degradation of accuracy (but not precision) of pseudorange and delta-range measurements. GPS users who are not authorized to use the GPS Precise Positioning Service (PPS) will not be able to benefit fully from performance benefits attributable to GPS/INS integration. Moreover, the two following constraints exist. First, the bandwidth of GPS code and carrier tracking loops cannot be reduced below the minimum required to track dithered GPS signals. This prevents the most aggressive use of INS aiding to reduce the dynamics tracked by the loops so that loop bandwidths can be reduced for purposes of increased radio-frequency interference rejection. Second, the complexity of optimum integration filters used to calibrate INS errors should increase to account for the artificial correlations among successive GPS measurements. Calibration will require longer observation intervals and will not converge as tightly compared to operations without the selective availability (SA) degradation.

The unauthorized user *can* still use GPS to generate position resets that keep the INS position errors below the SA limit, which is expected to be less that 100 m as measured by the $2d_{rms}$ criterion. However, in the absence of INS sensor *calibrations,* the INS error growth during a GPS outage will be faster than that for a calibrated INS. Assigning numbers to these qualitative comments requires either access to classified information or to measured values of SA waveforms obtained at well-surveyed fixed observation sites. As this information becomes available to civilian users, we should expect to see numerous contributions to the technical literature on this subject.

A. Operation During Outages

A stand-alone GPS receiver typically incorporates current measurements to four or more satellites to update its most recent PVT solution. Dead reckoning that incorporates recent estimates of vehicle acceleration may be used to propagate the current PVT solution in-between measurement updates. A GPS outage occurs when fewer than four valid satellite measurements are available at each update. During a partial or complete outage, the software for a stand-alone receiver can continue to produce a navigation output if it mechanizes one of the following options, albeit with reduced accuracy:

1) Compute the least-squares solution with fewer measurements than there are unknowns.[7]

2) Constrain one or more navigation outputs to be fixed, such as the UE clock bias or the vehicle altitude, or constrain the navigation solution to lie along a great-circle path.[8,9]

3) Incorporate measurements from an external sensor. A barometric indication of altitude is commonly available in military UE, as are radar- or pilot-inserted position updates.[9,10]

During an outage, the navigation solution becomes less accurate the longer the outage and the greater the vehicle dynamics since the last full set of measurements. The key factor to be specified when deciding whether an auxiliary sensor is required is the maximum acceptable error growth during the outage. In a conservative design, maximum error growth is calculated under worst-case conditions of vehicle dynamics.

Outages may be a concern even for UE that track more than four satellites at a time. For example, a GPS antenna mounted on top of an aircraft will only see a limited number of satellites during a banked turn, and the dilution of precision (DOP) parameters for that visible constellation may be unacceptably high. In more extreme cases, a vehicle passing through a tunnel may see no satellites for an extended period, and a military UE may be jammed as it approaches its target.

Combining GPS with an independent navigation sensor (item 3 in the preceding list) is one means to maintain the quality of the navigation service during a GPS outage. In effect, the independent sensor can act as a flywheel to provide continuous, high-quality navigation outputs. Inertial navigators are commonly considered for this role because they are passive, self-contained, and widely available. Moreover, they are not subject to the causes of GPS outage. However, they are generally more expensive to buy and integrate than other radionavigation sensors such as Loran or Omega. Their use has generally been limited to military and commercial aircraft. However, low-cost, low-performance inertial sensors implemented using mass-production microelectronics technology are emerging from research laboratories. These may provide the technological basis for an economical solution to GPS outages in general aviation and commercial applications, such as trucks and automobiles.[11,12]

With respect to GPS/INS integration performance during outages, the key questions are the following:

1) What quality INS is required?

2) How complex is the integration required to exploit the inherent INS quality to achieve mission objectives?

The resulting performance must then be weighed against the cost to determine whether to implement the optimum integration or to accept a less expensive, lower-performance solution.

B. Providing All Required Navigation Outputs

GPS UEs routinely estimate only the translational motion of a point referenced to the GPS antenna. Interferometric processing of GPS signals received at multiple antennas can also provide rotational (attitude) information.[13-15] However, we assert that an inertial solution is preferable to interferometry for terrestrial users whenever it is available. This preference is based on three considerations:

1) The inertial system is self-contained and is not vulnerable to outages (except those caused by equipment failure).

2) Installation of an inertial system on an aircraft is less demanding than an interferometer, and it is probably less demanding on ships and vehicles, also.

3) The noise floor on the accuracy of a short-baseline (1–5 m) interferometer has not yet been achieved. It seems that multipath is the culprit, and that it is premature to expect that antimultipath techniques will be effective and practical.[16]

In addition to attitude indication, the inertial navigator is desirable because its accelerometers typically sense velocity changes at up to a 1.0-KHz rate, with a 200-Hz output rate being commonly available. Therefore, the INS routinely outputs navigation solutions one to two orders of magnitude more often than a GPS UE. This high output rate allows the INS to provide accurate inputs to vehicle control subsystems, platform-stabilization systems, pilot displays, and velocity-aiding inputs to GPS tracking loops.

It follows that an integration in which GPS is used to bound the error growth of an INS-based system navigation solution would be very effective whenever GPS was available, and the availability of a calibrated INS may be the only means to maintain nearly as good performance during an outage.

C. Reduced Noise in GPS Navigation Solutions

In a stand-alone GPS receiver, the navigation processor usually implements a linear filtering algorithm in which the previous navigation solution is propagated to the current measurement epoch. Because GPS does not directly sense acceleration, the propagated solution is sensitive to errors in the previous acceleration estimate or to jerk that changes the true acceleration during the propagation interval. In contrast, an inertial system measures position change very precisely in the interval between GPS updates. This property can be exploited by a well-tuned Kalman filter, using GPS measurements to estimate *errors* in the INS output. Because these errors change slowly, the filter can smooth its update over many GPS measurements, thereby reducing the effect of additive noise on any one update. The result is that the "integrated navigation" solution seems to be much "smoother" than the stand-alone GPS solution. (See Sec. III.A, and especially compare Figs. 3 and 6 for an illustration of this feature.)

D. Increased Tolerance to Dynamics and Interference

The INS velocity solution may be fed back to the GPS UE to reduce the apparent dynamics of the input to the GPS code and carrier loops. This has two effects[1,4,6]:

1) A fixed bandwidth-aided tracking loop can maintain lock on GPS signals in the presence of dynamics that would cause the unaided receiver to break lock.

2) The tracking loop bandwidths can be reduced to the minimum amount required to track the *errors* in the INS aiding signals. (As noted previously, this feature breaks down for unauthorized users in the presence of selective availability clock dither.) (INS position errors are mostly low frequency.)

The net result of these actions is that the INS-aided GPS receiver can maintain lock and provide GPS measurements over a much wider range of vehicle dynamics and radio frequency interference than the unaided, stand-alone receiver.

II. GPS Integration Architectures and Algorithms

The degree of complexity of the integration should reflect the mission requirements; it may also be limited by the investment that can be made to obtain those objectives. Integration strategies and mechanisms may be very simple (for example: choose the GPS UE position and velocity as the integrated solution when GPS is available with a given precision, otherwise choose INS position and velocity as the integrated solution) or relatively complex (for example: optimally combine GPS UE measurements with INS outputs, Doppler radar outputs, baroaltimeter signals, true airspeed, and other sensor data). However, in the following remarks, we limit our attention to alternatives involving only GPS integrated with an inertial system.

A. Integration Architectures

Figure 1 illustrates three generic functional architectures for GPS INS integration. The GPS receiver and the INS are treated as navigation *systems* in architectures a and b, with GPS supplying a position, velocity, and time solution, and the INS supplying a position, velocity, and attitude (P, V, θ) solution, respectively. In architecture c, the GPS and INS are treated as sensors producing line-of-sight measurements ($\rho, \dot{\rho}$) and accelerations and angular rates ($\Delta V, \Delta\theta$), respectively. In addition to the GPS and INS units, each architecture includes various data paths and a processor unit that mechanizes the integration algorithm. These alternatives are distinguished by the data passed between the subsystem components. The proper interfacing and control of these components may incur the largest part of the cost of an integration project, but those concerns are not within the scope of architectural considerations.

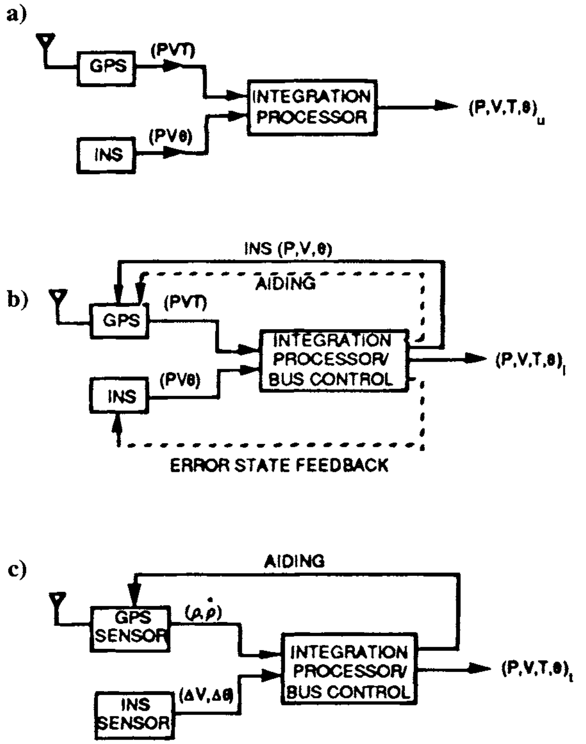


Fig. 1 Generic global positioning system/inertial navigation systems architectures: a) uncoupled mode; b) loosely coupled mode; c) tightly coupled mode.

1. Uncoupled Mode

Figure 1a illustrates the configuration in which GPS UE and an INS produce independent navigation solutions with no influence of one on the other. The integrated navigation solution is mechanized by an external integration processor that may be as simple as a selector or as complex as a multimode Kalman filter. All data busses are “simplex” (unidirectional). The characterization of Fig. 1a as an “uncoupled” mode is based on the independence of the GPS and INS navigation functions. Note that, in principle, the *hardware* could all be packaged in one physically integrated (embedded) unit; however, the functionality would still be that of uncoupled architecture.

The potential benefits of integrating the navigation solutions from uncoupled GPS and inertial navigators are:

1) It is the easiest, fastest, and potentially the cheapest approach when an INS and GPS are both available.

2) It provides some tolerance to failures of subsystem components (except in the embedded configuration, see Sec. II.C).

3) Using an integration processor as simple as a selection algorithm can provide en route navigation at least as accurate as available from an INS.

2. *Loosely Coupled Mode*

Figure 1b illustrates a configuration in which there are several data paths between the integration processor and the GPS and the INS equipment. Among these, the provision of the system navigation solution to the GPS UE is the most important for getting the maximum benefit from the integration filter. The inertial aiding of GPS tracking loops is of next greatest benefit, and feedback of error states to the INS is of second-order benefit. There may also be some improvement of system reliability to the extent that individual components are likely to be more mature and to have been better tested than emerging technology that features more highly integrated subsystems.

a. Reference navigation solution. GPS UE generally employs a Kalman filter mechanization to compute PVT updates based on current tracking loop measurements. A GPS UE does not directly sense acceleration; it must use relatively noisy acceleration estimates based on recent velocity measurements for a dead reckoning propagation of the previous navigation solution forward to the epoch of the current tracking-loop outputs. The situation changes dramatically when the system navigation solution is fed back to perform that propagation. In effect, the GPS measurements can now be used (within the UE navigation filter) to correct the system navigation solution. Over short periods of time, that solution is very accurate because it incorporates INS data based on acceleration sensing. The UE filter is then mechanized to estimate INS (or system) error states having relatively low dynamics and low uncertainty (process noise). The filter can be tuned to have a longer time constant (filter memory), thereby increasing the effective averaging of each noisy GPS measurement.

b. Inertial aiding of GPS tracking loops. As mentioned in Sec. I, the availability of a GPS navigation solution can be increased significantly when inertial aiding is used to reduce the vehicle dynamics tracked by the UE code and carrier loops. In principle, this aiding could be applied directly from the INS to the GPS UE, but it is shown as an output of the integration processor in Fig. 1b because of the following:

1) GPS tracking loops must be aided by the projection of vehicle velocity along the line-of-sight (LOS) to each satellite being tracked. The conversion from inertial coordinates to GPS LOS coordinates is most appropriately done in the integration processor or in the GPS UE itself. In either case, INS velocity information is available within the processor hence aiding can be part of the data flow to the UE. This avoids the expense and risk of developing a custom interface from the INS to the GPS UE.

2) Executing the coordinate transformation external to the INS retains flexibility in the selection of INS equipment and avoids the need to develop custom GPS/INS interfaces for each application. However, this raises a concern for "data latency" (i.e. feeding delayed data to the tracking loops) as mentioned in Sec. II.C.

c. Error-state feedback to the inertial navigation system. Most inertial navigation systems have the means to accept external inputs to reset their position and velocity solutions and to adjust the alignment of their stable platform. The adjustment may be executed by a mathematical correction in a "strap-down" inertial system, or it may be realized by torquing a gimballed platform. In either case, the use of feedback can maintain inertial navigation errors at a level for which their dynamics are accurately modeled by the error state propagation equations embodied in the integration filter. However, the impact of this feedback on error growth for a navigation grade INS is relatively small until the errors grow much larger than 10 km and 1 m/s, respectively.

3. Tightly Coupled Mode

Figure 1c illustrates the so-called tightly coupled integration mode. It differs from the loosely coupled mode in that both the GPS receiver and the inertial components are limited to their sensor functions. They are treated as sources of GPS code and carrier measurements and inertial indications of acceleration (velocity change) and angular rate, respectively. These sensor outputs are then combined in one navigation processor that may mechanize an appropriately high-order integration filter.[17-19]

In the tightly coupled mode, there is only one feedback from the navigation processor. Figure 1c illustrates the use of velocity aiding to the GPS tracking loops. Acceleration aiding could also be effectively used, but we are not aware of any particular mechanization using other than velocity aiding. The other paths used in loosely coupled architectures are not needed here because all computations involved in navigation processing are now internal to one processor.

The concept of tightly coupled integration is often raised in connection with embedded GPS receivers. These are not necessarily synonymous. However, it is reasonable that we would choose to mechanize a tightly coupled integration algorithm if we had already taken the effort to design a GPS receiver that is physically and electrically integrated with an inertial sensor or with a powerful navigation processor. We return to this point in Sec. II.C.

B. Integration Algorithms

The basic choices for GPS integration algorithms are 1) selection, with or without INS resets; 2) fixed-gain filter; and 3) time-varying filter. These are listed in order of increasing complexity and optimality. Each one can be used with any one of the architectures listed in Fig. 1, but the incremental payoff of a more complex filter is directly related to the quality of the input information.

1. Selection

A selection algorithm chooses the GPS indicated (PVT) as the system navigation solution whenever the GPS UE indicates that this solution is within acceptable bounds on its accuracy [via the GPS figure-of-merit (FOM)]. Inertial navigation systems data may be used to interpolate between successive GPS updates when a higher output rate is needed than can be provided by the UE. During GPS outages, the INS solution extrapolates from the last valid GPS solution. (The process of forcing the INS solution to equal the current GPS indicated velocity and/or position is known as a "reset" if that correction is actually fed back to the INS.)

2. *Filtering*

The general filtering problem involves trying to estimate time-varying states whose evolution is characterized by known laws of propagation, which usually are taken to be a coupled system of linear differential equations driven by white noise.

States usually cannot be measured directly, but they are inferred from measurable quantities to which they are related. These measurements may be made simultaneously or sequentially at a series of distinct points in time. The filter will generally incorporate knowledge of the statistics of the measurements.

Knowledge of the way the states change (propagate) in time, knowledge of the way the measurements are related to the states, measurement statistics, and measurement data are all used in each state update. The most common update algorithms use linear filters; e.g., ones in which the updated state is a linearly weighted sum of the measurements and the previous state value.

Position and velocity of an aircraft are examples of quantities that may be chosen as *states* in a filter (these are referred to as whole-value filter states). For whole-value position and velocity states, the propagation equations are simply the equations of motion of the aircraft. To make the whole-value filter propagation equations a better reflection of the real world, acceleration states could be added (otherwise, by its omission, acceleration must be treated as "noise," driving the derivative of velocity). GPS-indicated position and velocity are examples of *measurements* that might be processed by an integration filter with whole-value states. At one extreme, the integration filter could ignore everything except the GPS receiver position and use this as the integrated position. This degenerate case is the selection mode cited above in which the state propagation equations and any other available measurements would be ignored. For the degenerate case, the weight on the GPS UE position is one and the weight on the propagated state is zero. The weight on the measurement is referred to as the filter *gain*. In general, some rule must be used in order to determine how much weight should be put on a measurement and how much weight should be put on the propagated states.

Another choice of states are the *errors* in position and velocity indicated by the INS (these are referred to as error states). For a filter whose states are INS errors, accurate representations and linear approximations of the propagation equations are well known. As in the case of whole-value states, additional INS error states (for example, states for azimuth and tilt errors, accelerometer bias, and gyro drift) could be added to the filter in order to make the propagation equations a better model of the real world. Of course, the degree to which the filter must reflect the real world is a function of the estimation accuracy required, and that is a reflection of the mission requirements.

For a GPS/INS integration filter with INS error states, the measurements would actually be the differences between GPS position and INS position and the differences between GPS velocity and INS velocity. As with the case of whole-value states, some rule must be used in order to determine how much gain should be put on the measurements and how much weight should be put on the propagated states when computing state updates.

We should mention in passing that the optimum filter may require an impractically large number of states. Options to decompose a high-dimensional estimator into combinations of lower-dimensional filters have been described in the literature. Distributed filtering and federated filtering are the terms under which these options are usually cited, as in Refs. 20 and 21. Both are believed to be more robust than the optimum filter when the design must be tolerant to imperfect information about the estimation problem, and it is claimed that federated filters are more fault-tolerant. These details are beyond the scope of this survey.

a. Fixed-gain filters. In a fixed-gain filter, the propagated estimates are combined with new measurement data using predetermined gains. The gains are fixed in the sense that they have been loaded into computer memory a priori, so that the filter selects from a short list of gains, rather than computing them. Different gains may be used for different sensor status and operational status, reflecting the uncertainties in the propagated solution and in the measurements. In general, the gains in a fixed-gain filter can have any value (they should at least properly reflect the relationships among the measurements and the states).

If the state dynamics and their uncertainty are limited, and there is negligible variation of measurement noise during the interval of interest, it may be that the optimum filter gains will not vary much during the mission. In that case, the performance penalty of mechanizing one fixed set of gains (or a few selectable sets of gains) compared to optimum time-varying gains may be acceptably small. The benefit to the integrator is a vast decrease in computational burden and memory required to implement the filter. It may even be effective to precalculate an approximation to time-varying Kalman gains that can be stored for use during a mission.

b. Time-varying gain (Kalman filter). In the Kalman filter, new gains are computed every time measurements are available. The Kalman filter is a recursive implementation of the optimum least-squares error estimation algorithm. It is optimum in the sense that it strikes the correct balance between uncertainly in the presumed dynamics of the states being estimated (process noise), uncertainly in the measurements (measurement noise), and the observability of individual states (sensitivity) required to minimize the figure-of-merit. See Refs. 22–24 for a detailed discussion of Kalman filtering. In the present context, we note that the updating of N states by M measurements involves substantial matrix manipulations, propagation of difference equations, and memory to store the matrices. Current technology can handle updates of around 20 states at up to a few times per second in a reasonably cost-effective processor. Because upwards of 100 error sources may influence an integrated GPS INS solution, the brute force approach to real-time integration is not yet computationally feasible. Each designer must complete detailed design studies to determine the minimum number of states and the update rate that will result in an acceptable navigation error using the available processor resources and with acceptable design margin. Given the rapidly changing computational capabilities available to avionics integrators, questions of computational feasibility should be reconsidered every few years.

3. *Discussion*

The uncoupled mode is inferior (in performance) to the loosely coupled integration mode. The uncoupled and loosely coupled integration modes are inferior (in performance) to a tightly coupled mode because information inherent in the sensor measurements is lost in the receivers mechanization of the PVT solution; i.e., it is not always possible to backtrack from a PVT solution to the raw GPS measurements with sufficient bandwidth and precision to support a tightly coupled integration. The feedback of the system navigation solution to the INS (via resets) is a second-order improvement in the loosely coupled mode (and is inherent in the tightly coupled mode).

The current generation of military high-dynamics GPS UE (receiver 3A, MAGR) mechanizes an integration filter when it is operated in its "INS mode." This internal filter is tightly-coupled in the sense of Fig. 1c. However, its performance is suboptimal because it incorporates a very simplified model for the dynamics of its inertial error states, and because the filter is tuned very conservatively. Integrators who need better performance usually resort to a higher-order external integration filter that combines the RCVR-3A navigation output with the inertial navigation solution. This cascaded integration (e.g., filter-driving-filter) is often criticized by proponents of tightly-coupled integration. This criticism is valid, but it really addresses a cost tradeoff in which the integration filter in RCVR 3A was limited to 12 states, of which 9 represent very generic INS errors (P, V, θ). The decision for RCVR 3A was based on unit cost and the desire to produce a generic standard equipment that did not burden any user with features not justifiable in his or her application. In principle, a modified RCVR 3A operating in the INS mode with an expanded internal filter and appropriate software could perform as well as an externally mechanized tightly coupled integration.

In design studies of tightly coupled GPS/INS integration filters, as many as 80 inertial error states are modeled, in addition to GPS error states related to delay measurement bias, tracking loop errors, propagation errors, and user clock errors. In some ultraprecise systems, it may even be useful to incorporate additional states that model multipath effects. Nevertheless, many studies have shown that most of the benefit of expanded error state formulations is gained with 25–30 states, and that adequate performance can usually be obtained from 14–17 states.[25]

The tightly coupled mechanization does avoid one problem commonly attributed to loosely coupled integration, namely the possibility of instability (in state estimates) arising when the GPS navigation errors become highly correlated with INS navigation errors. This situation may occur at low input signal-to-noise ratios when GPS code loops remain in lock only because inertial aiding allows the loop bandwidth to be reduced, thereby reducing the effective levels of noise and interference. Now, the narrower the loop bandwidth, the more the loop error approximates the error of the aiding signals so that the correlation cited above becomes significant. See Ref. 26 for further discussion of this point.

C. Embedded Systems

As GPS approaches its operational status, there has been a massive increase in investment in civil GPS technology, which has led to smaller, lower power-consuming, higher-performance UE than were dreamed of as recently as the late 1980s. One consequence of this trend is that GPS UE can be packaged on a single card that can be embedded in other systems. As noted in Sec. II.B, the concept of GPS embedded in an INS is one such application that is being prominently discussed at present, with several efforts underway to demonstrate the concept.[27–30]

Setting aside the valid claims of savings in size, power, and weight that accrue from embedding, it is reasonable to ask whether there is any functional or performance payoff directly attributable to embedding. The answer is a qualified yes. There are potential performance improvements, but the system may be vulnerable to a single-point failure, such as a power supply or the processor.

1. *Tight Coupling*

There is no inherent reason to claim that embedding implies tightly coupled integration. An embedded receiver could be stand-alone, loosely coupled, or tightly coupled. However, developers of embedded systems have tended to mechanize tight coupling as a performance feature.

2. *Carrier Loop Aiding*

Standard military UE use inertial aiding only for code loops and only after carrier loops have lost lock. The decision to limit the INS aiding goes back to the late 1970s when it was argued that the latency (time delay) between the sensing of inertial velocity and its receipt at a GPS receiver could be as large as tens of milliseconds, even with the high-speed data busses that were available. With this much delay, it was argued that errors in the aiding signal during accelerations or turns could be large enough to drive the carrier loop out of lock.

There are at least two ways to mitigate this concern in an embedded system. The most common approach is to customize the data link between the INS and the GPS carrier loop to reduce the latency to a few tens of microseconds and to minimize the uncertainty in the latency. With that small a delay, the maximum error of the aiding signal is negligible, even for an aircraft rolling as fast as 1 rps, and moving toward a satellite with a relative velocity of 2000 ft/s. Under those conditions, the error caused by a 20-μs delay in attitude indication would be approximately V_ϵ, where

$$V_\epsilon < 2\pi\ (20 \times 10^{-6})\ \text{rad} \times 2000\ \text{ft/s} < 0.25\ \text{ft/s}$$

which is well within the acceptable range for GPS receivers. An alternative that has not yet been mechanized is to delay the GPS signals by an amount that matches the latency, before the aiding signal is applied. For modern precorrelation digital GPS receivers, we could store tens of milliseconds of GPS samples in a data buffer mechanized by a single memory chip.

3. *Tracking Fewer Than Four GPS Satellites*

The loose-coupling approach (Fig. 1b) integrates a GPS PVT solution with an inertial (P, V, θ) solution. When the GPS solution is unavailable, the integrated solution "flywheels" using the inertial solution as corrected at the start of the

GPS outage. In contrast, the tightly coupled solution uses raw GPS measurements, which are available as long as one or more satellites is being tracked. Thus, it is a more robust solution vs outages that could prevent a GPS navigation solution from being formed in the loosely coupled configuration. Reference 25 gives a good insight into the potential performance improvement (see also Sec. III).

However, there is one caveat. Conventional GPS UE, such as GPS RCVR 3A, can continue to provide a navigation solution (albeit degraded) with only two or three satellites. Thus, it is inappropriate to claim that a loosely coupled integration *must* convert to a free-running inertial solution in the presence of one or more satellite outages. The performance will depend on the details of the GPS UE mechanization (see Sec. III).

4. Quantization

All calculations within an embedded system are more likely to be executed as "full-precision" quantities than in a system wherein the GPS and INS and navigation processor are connected by data busses. These busses are usually so heavily used that data must be coarsely quantized for data transmission (compared to their internal precision) in order to satisfy communication bandwidth constraints and to conform to data transmission protocols. An alternative that has not been explored in GPS navigation data communications is to use data compression to increase the information content of the message structure. This would require reworking interfaces and message protocols, but the effort might be cost effective in high-precision applications.

III. Integration Case Studies

We consider three case studies that illustrate performance benefits of GPS/INS integration. The properties to be addressed include 1) noise quieting (reduced variance) with an INS error state filter; 2) in-flight INS alignment; and 3) reduced error growth of a "calibrated" INS during a GPS outage.

A. GPS/Inertial Navigation Systems Navigation Performance in a Low-Dynamics Aircraft

In 1985, a five-channel GPS UE (RCVR-3A), operating as a stand-alone system, was flown in a DeHaviland Twin-Otter aircraft that was also equipped with a laser-inertial integrated navigation system. The combination of this very stable inertial system with precise angle-angle-range pulsed optical measurements to surveyed [by the U.S. Geodetic Survey (USGS)] retroreflectors removed the long-term increase of inertial position errors attributable to drift, misalignment, and gravitational anomalies, and tied the reference solution to local geodetic coordinates.[31]

With postprocessing, the laser-inertial reference system located the aircraft to within 50 cm (position) and 5 mm/s (velocity) at any time during the flight, and to within 1.0 cm (position) and 0.3 mm/s during lock-on to a retroreflector. Therefore, it was at least one order of magnitude more accurate than GPS UE specifications and was, therefore, uniquely suited to score the inflight performance of the GPS navigation system. It was the most accurate reference system ever used in GPS flight testing.

The GPS UE was operated in the stand-alone PVA mode; however, each pseudorange and delta-range measurement was recorded in addition to the navigated UE solution and supporting data. Therefore, the performance of the receiver with different navigation processing algorithms could be evaluated by postflight emulations, and then compared to the stand-alone output and to the reference solution.

During the flight tests, the aircraft flew six circuits around five retroreflectors placed in suburban Boston (Fig. 2). The reference solution (altitude, latitude, longitude) is shown in Fig. 3.

1. GPS Point Positioning

The GPS pseudorange measurements were processed to form a point-positioning solution. In other words, the navigation equations were solved for each set of GPS measurements as a single set of four equations with four unknowns. In this case, there is no navigation filter; nor is any "memory" of previous navigation solutions used to smooth the results. If the noises on each satellite measurement were the same, then the position errors would be proportional to the position dilution of precision (PDOP) for the collection of satellites being tracked.

Figure 4 illustrates the resulting navigation errors, and Fig. 5 illustrates the DOP values that were current during the data collection. (In regions where Fig. 4 is multivalued, the smaller value is the minimum DOP, and the larger value corresponds to the DOP for the satellites actually selected.) The discontinuities

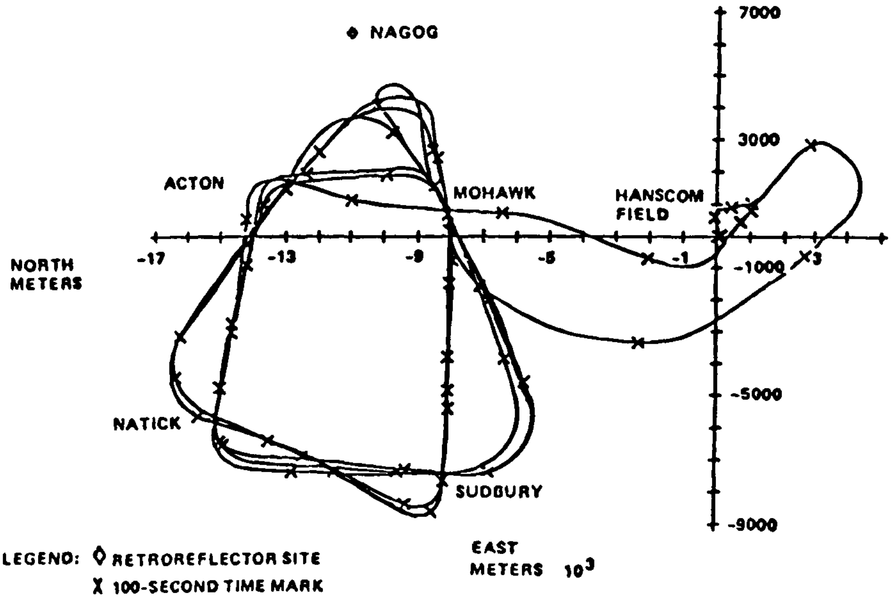


Fig. 2 Flight path on May 30, 1985.

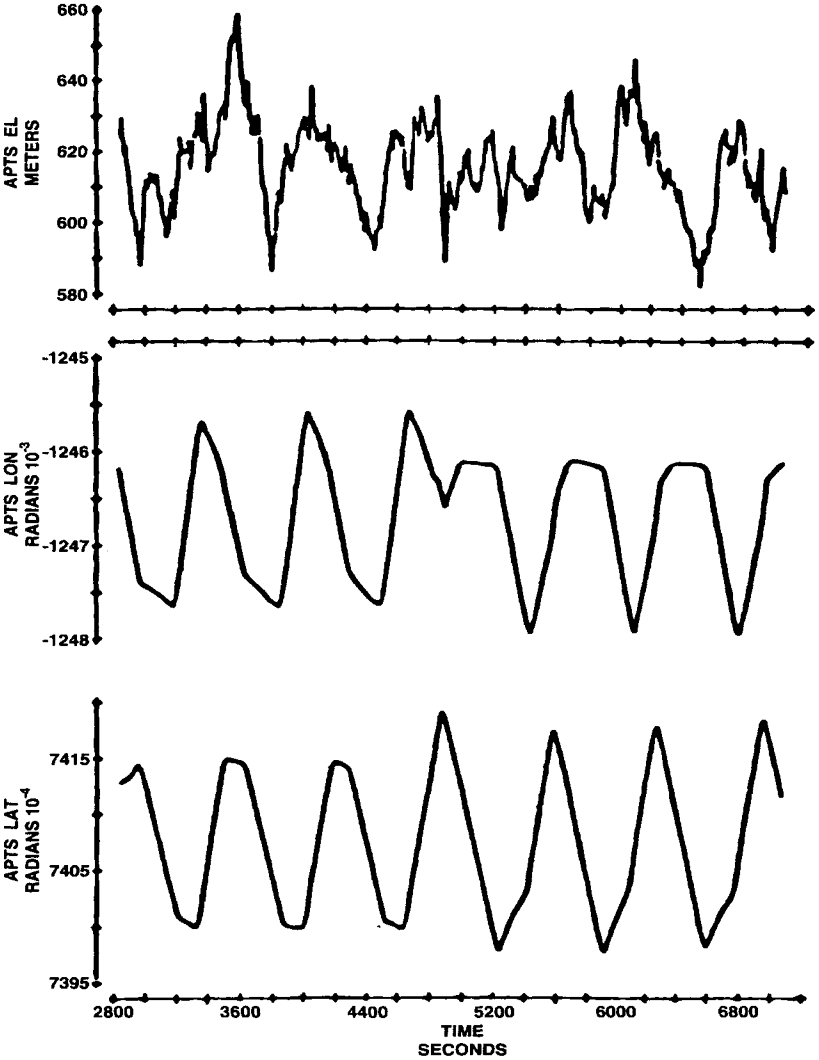


Fig. 3 Reference navigated position on May 30, 1985.

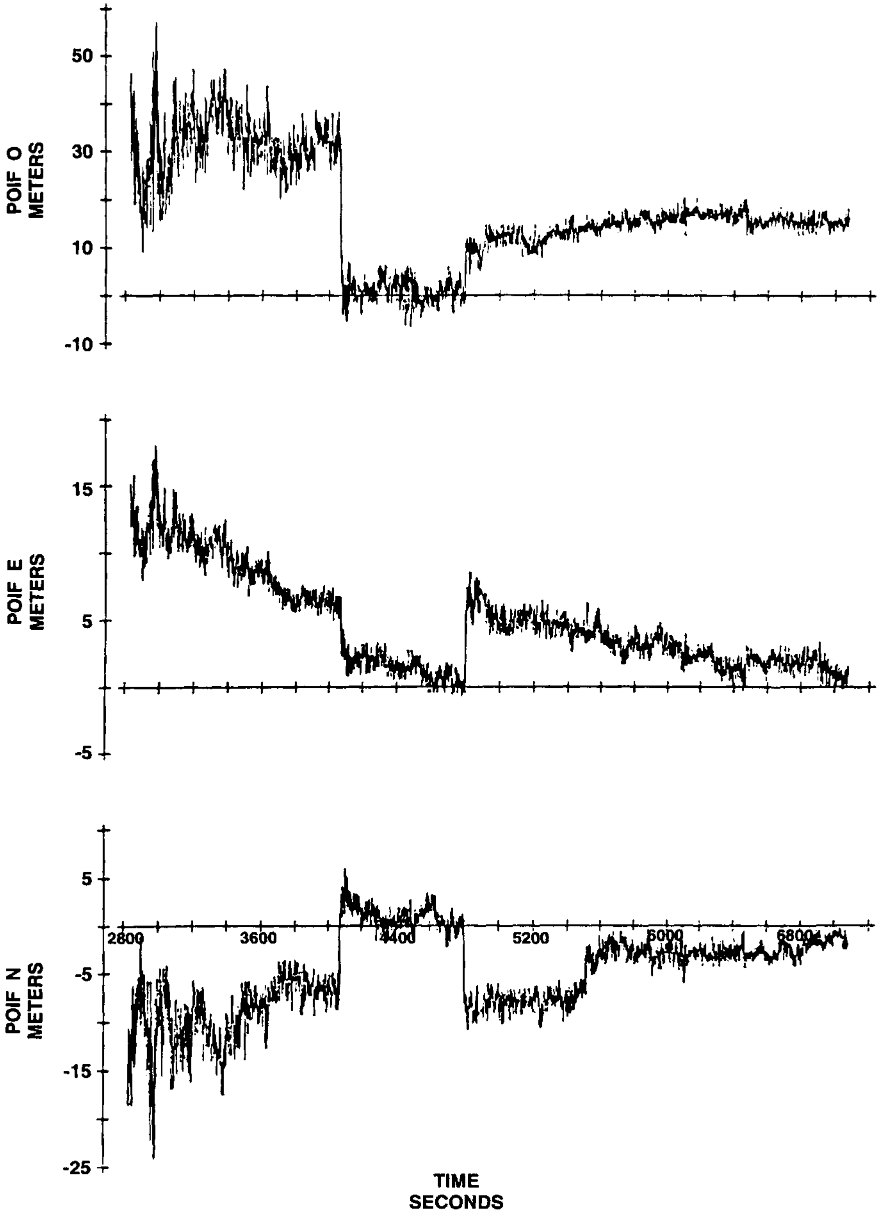


Fig. 4 Navigation errors (position) for point positioning with unaided five-channel receiver.

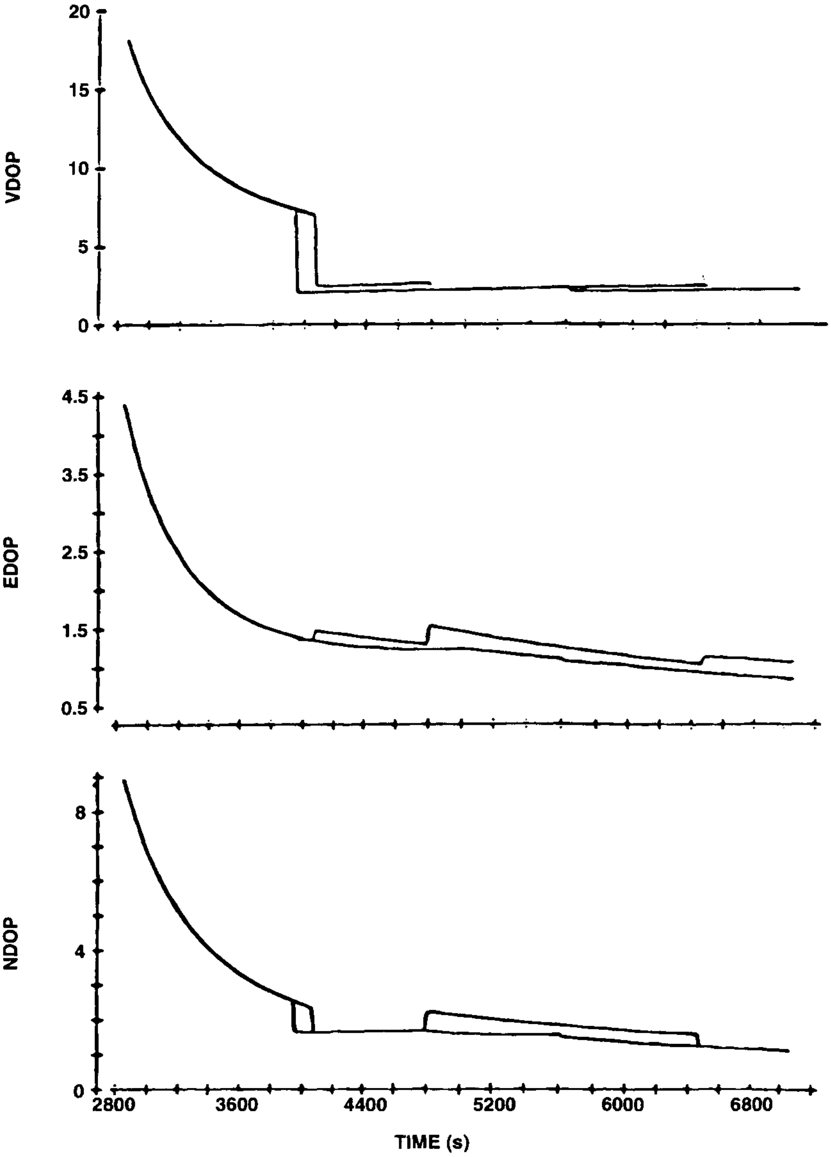


Fig. 5 Navigation dilution of position, EDOP, and vertical dilution of position on May 30, 1985.

in DOP signify satellite switches. The hypothesis that the point-positioning errors are proportional to DOP is supported by a comparison of these two figures.

2. *GPS Internal Navigation Filter*

Figure 6 illustrates the errors in the position solution output by the receiver, using its internal 12-state PVA-mode navigation filter. The errors are the difference between the RCVR 3A solution and the laser-inertial reference solution. The RCVR 3A navigation filter implements 12 error states including 9 for three-dimensional position, velocity, and acceleration (hence, the name PVA mode), and 2 for the internal receiver clock drift (seconds), and clock drift rate (frequency) offset from their nominal values. The 12th state calibrates bias in barometric altitude when that measurement is available. The baro output was not used in these flight tests, so only 11 states were updated by the filter.

Comparison of Figs. 4 and 6 show some of the features of the filtered solution compared to the point solution:

1) The influence of satellite constellation changes (e.g., at about 4100 s and 4800 s) is substantially reduced.

2) The noise standard deviation (estimated as one-half of the peak to peak variation of short-term errors) after "good" geometry was established (about $t = 4100$ s) is reduced by filtering from 2.5 m to about 1.0 m per coordinate.

3) Errors in the filtered solution show smaller values of bias, and they are more nearly a zero-mean process. We speculate that the reduction in bias errors is a consequence of the particular "tuning" of the receiver filter, which causes the biases to show up more in the clock error states than in the position error states.

Further analysis of the field test data showed large error spikes in the GPS velocity solution at each turn. These errors were consistent with an apparent lag in the GPS output relative to the reference solution. Indeed, it turns out that the RCVR 3A PVA filter output epoch is the *end* of the delta-range accumulation interval. This is 0.39 s later than the *middle* of that period, which is the effective epoch of the delta-range observable. This underscores the need for an integrator to be aware of all potential sources of time bias between GPS, inertial, and other sensors that are being integrated.

3. *GPS Navigation in the Inertial Navigation Systems Mode*

Inasmuch as the pseudorange and delta-range measurements from the receiver were available, as well as all the inertial outputs, it was possible to emulate the navigation solution that *would* have been produced by RCVR 3A operating in the INS mode. With respect to that particular receiver, we note that this procedure is valid because:

1) The receiver tracked in "State 5" throughout the flight tests. Therefore INS "aiding" would not have been applied to the receiver tracking loops, even if the receiver had been operated in the INS mode during the flight tests.

2) The bandwidth of the tracking loops is the same as would have been in effect in the INS mode. Therefore, the measurement noises input to the navigation filter emulation have the same standard deviation as the measurement noises for operation in the INS mode.

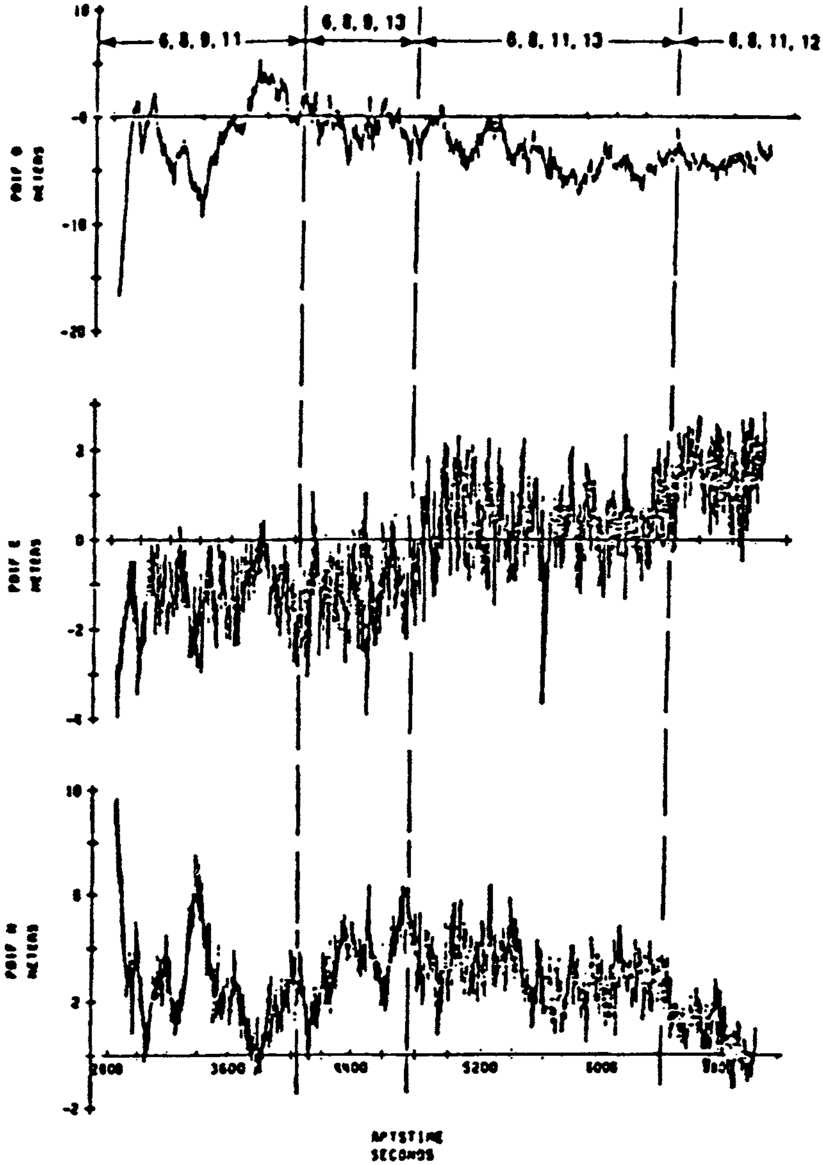


Fig. 6 Difference between global positioning system and laser-inertial navigated position on May 30, 1985.

Figure 7 illustrates the navigation solution output by a 17-state Kalman filter with measurements available once per second. The seventeen states are: position error (three states); velocity error (three states); UE clock error (two states); inertial platform misalignment (three states); gyro drift rate (three states); and accelerometer bias (three states). The 17-state filter adds 6 error states for gyro drift rate and accelerometer bias that are not available in the RCVR 3A navigation filter. Thus, the results are more suggestive of performance to be obtained with certain tightly coupled GPS/INS integrations than of the RCVR 3A performance.

The most striking feature of Fig. 7 is the suppression of additive noise. The navigation errors track satellite DOP variations with greatly reduced noise compared to the point solution shown in Fig. 4. The position errors are clearly dominated by slowly varying biases in the pseudorange measurements to each satellite. A secondary feature is the treatment of errors in the period from t = 2800 s to t = 4100 s when the satellite geometry is poor. The 17-state filter is clearly "tuned" differently than the 11-state filter in RCVR 3A. (Filter tuning is the process of adjusting filter parameters that model measurement noise and process noise in the computation of the filter weights at each update.) This tuning is probably closer to optimum because the pseudorange and delta-range residuals computed using those clock error states, were smaller than residuals computed using the receiver's clock error states. This example suggests the influence that filter tuning can have on navigation performance.[32]

Finally, we note that for "Navigation grade" inertial systems (commonly defined as having 1.0 nautical mph error growth attributable to gyro drift), the error growth of a GPS "calibrated" unit is dominated by the initial velocity error, accelerometer bias, and gravity anomalies for about the first 10 min after a GPS outage. Gyro bias begins to dominate around 20 min (about 0.25 Schuler period). The effectiveness of any "calibrations" on reducing errors at 20-min, or longer, intervals depends on the level of "random walk" gyro and accelerometer errors that accumulate in that interval.

B. Using GPS for In-flight Alignment

Alignment is the process that ties inertial platform coordinates to the geographic frame in which the host vehicle navigates. Alignment establishes the conditions necessary for the proper integration of accelerometer outputs to accurately estimate the changes in user position and velocity.

Gyrocompassing typically extends over a 10–15-min period prior to a standard aircraft takeoff. We need to look at the factors that contribute to alignment time in order to understand the opportunities for using GPS to speed up the process, or to allow it to proceed in-air in addition to on the ground.

1. Conventional Alignment Procedures

Alignment consists of platform leveling and establishing a reference bearing. The following discussion is based on the alignment of a gimballed platform INS, but the general conclusions and timing estimates are good approximations to the performance of a "strapdown" INS.

a. Leveling. Leveling is the process of establishing one plane of the inertial measurement unit (IMU) instrument package perpendicular to the local gravity

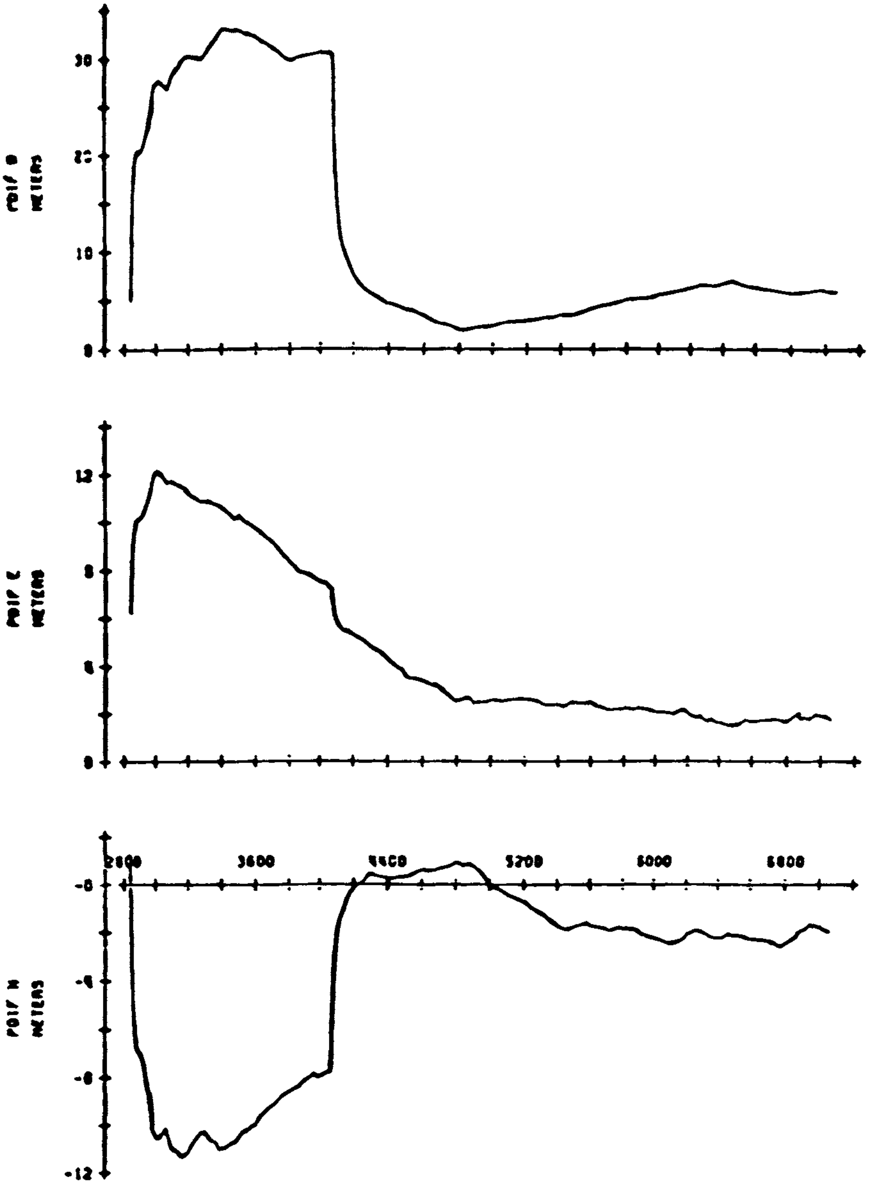


Fig. 7 Position errors with aided five-channel receiver and 17-state Kalman filter.

vector. This is typically accomplished by mounting two accelerometers on that plane so that their input axes are not co-linear (they are usually mounted at right angles), and tilting the plane to null the accelerometer outputs. (Alignment for "strap-down" inertial systems is a mathematical operation that does not involve physical motion of the sensor platform.) The perpendicular to this plane defines "UP" in an East, North, Up coordinate system.

The error signal that drives the leveling loop is proportional to the instantaneous value of the tilt error; i.e., to approximately 1.0 g times the alignment error (in rads). In mathematical notation, the output for each horizontal accelerometer is $\hat{g}$, where

$$\hat{g} = (1 + s_f)g\, e_\Phi + b$$

where g is the magnitude of the local gravity vector; e_Φ is the angular error of the platform normal to the accelerometer axis (i.e., the tilt error); b is the accelerometer bias; and s_f is the accelerometer scale factor error. For an aircraft at rest, $\hat{g}$ should be zero when the platform is level; however, the accelerometer bias causes the null to occur for a nonzero value of tilt error. For accelerometers used in contemporary fighter aircraft, b is about 150 μg, and s_f is about 500 ppm. For $g = 1.0$, the steady state tilt error e_Φ is then approximately

$$|e_\Phi| = b/(1 + s_f) \approx 150\ \mu\text{rad} = 30\ \text{arcsec} \tag{1}$$

where we assume that angle quantization effects in the leveling feedback loop are negligible. This 30 arcsec is a bias error; the random component of leveling errors is on the order of a few arcseconds.

b. North seeking. North seeking is the process of establishing a reference direction (azimuth) in the leveled plane containing the east and north accelerometers. The most widely used scheme for self-alignment of the platform is gyrocompassing.

Gyrocompassing exploits the following properties of gyroscopes. For a local level north-oriented system at latitude λ

1) If the gyro input axes are physically aligned with the geographic axes (North, East, Up), then the system will remain aligned if each gyro is individually torqued at a rate equal to the projection of the Earth rate on its input axis.

2) If the platform is level, but not aligned to north, it will rotate around the east axis, causing the north axis to rotate from horizontal. This rotation produces a level error than can be sensed by the north accelerometer and used to drive the azimuth error to zero.

The rate of rotation of the north axis in response to an azimuth misalignment of θ_A radians is given by $\dot{\theta}_N$

$$\dot{\theta}_N = \theta_A \Omega_e \cos\lambda$$

where Ω_e is the Earth rate. If this rate acts for t seconds, it produces net level error (tilt) of θ_N (radians)

$$\theta_N = \theta_A \Omega_e\, t \cos\lambda \tag{2}$$

From Eq. (2) we see that for gyrocompassing to convert a misalignment error

θ_A into an equally large observable tilt error (i.e., for $\theta_N = \theta_A$), then the effect of the error must be integrated for at least t_o seconds, where

$$t_o = (\Omega_e \cos \lambda)^{-1}$$

At midlatitudes, (say $\lambda = 45$ deg) and for

$$\Omega_e = 2\Pi \text{ rad/day}$$

we have

$$t_o = \frac{1}{\sqrt{2}\Pi} \text{ days} = 5.4 \text{ h} \tag{3}$$

In practice, the requirement for alignment accuracy is substantially less than the requirement on leveling accuracy, and this allows the alignment to proceed faster than indicated by Eq. (3). We can estimate this speedup using the approximation that the standard deviation of leveling errors is inversely proportional to the time spent in leveling. For medium accuracy inertial sets, which are found in contemporary aircraft, the random noise in leveling is on the order of a few arcsec, and the uncertainty in azimuth alignment after gyrocompassing is on the order of 160–200 arcsec. The ratio of these standard deviations is about 50. Substituting $\theta_A = 50\ \theta_N$ in Eq. (2) yields $t_o = t_o/50$, which is approximately 7 min. This estimate is consistent with Air Force specifications for the time to achieve standard accuracy alignment by gyrocompassing under favorable conditions.[33]

2. *GPS Aided Alignment*

Leveling is essentially an instantaneous process (especially in a strap-down system) because it seeks to drive two directly sensed gravity indications to be equal. This gives the clue as to the benefits of inflight alignment. Velocity changes sensed by the inertial system are compared to velocity changes sensed by GPS, and the alignment parameters are adjusted to drive the residuals to zero. Sensing INS alignment errors from velocity residuals rather than from position residuals (as in traditional gyrocompassing) is significant because it eliminates the delay incurred to integrate a velocity error to a position error of detectable magnitude. Now, combine this observability with the opportunity to use aircraft maneuvers (turns, climbs, dives) to apply acceleration in the horizontal navigation plane (North, East), in addition to the vertical plane (Up), and the potential use of GPS for inflight alignment is clear.

The issues for in-flight INS alignment using GPS are related to the noisiness of GPS velocity measurements compared to inertial measurements. Unless the INS misalignment is so large that, at the vehicle velocity, it produces a velocity error that exceeds the GPS noise level, the alignment will require several minutes, or more, of observations. Thus, the time to align is a function of the accuracy goals, the magnitude of acceleration that can be applied during maneuvers, and the noise level of GPS measurements.

Figures 8–10 were generated in a linear covariance simulation of INS alignment for a fighter aircraft using a standard (1.0 n.mi./h) inertial navigator. The integration filter mechanized 11 states (3 position, 3 velocity, 3 misalignment, 2 clock).

The measurements input to the filter are GPS pseudorange and delta-range and inertially indicated velocity change projected along the nominal line-of-sight toward each satellite being tracked. Although the filter is similar to the internal filter in RCVR 3A, its performance is superior because the inertial platform dynamics are more accurately modeled and the filter tuning (process–noise values) reflects the nominal INS error variances rather than worst-case values used in the RCVR 3A tuning. The results presented can be viewed as an upper bound on the performance available from RCVR 3A when the filter accurately models the integration environment. The RCVR 3A filter trades suboptimum performance for reduced sensitivity to any mismatch between actual INS performance and the INS as modeled in the filter. Two takeoff profiles were investigated.

a. Standard takeoff. 1) accelerate at 0.3 *g* to 160 knots indicated airspeed (KIAS) using MIL (standard) acceleration; 2) begin rotation to 5-deg pitch angle and accelerate to 170 KIAS takeoff speed; 3) climb at 5 deg, and accelerate to 385 KIAS, on 45-deg heading (northeast); and 4) Continue climbing at 5 deg and constant velocity to cruise altitude at 20,000 ft, and 385 KIAS.

b. Alert (scramble) takeoff. 1) accelerate at 0.4 *g* to 165 KIAS using maximum afterburner thrust; 2) rotate to 12-deg pitch angle and climb at 0.4 *g* to 415 kt at 2500 ft altitude; 3) climb from 2,500 to 20,000 ft altitude at constant speed and 12-deg pitch, arriving at 20,000 ft approximately 180 s after takeoff; and 4) level flight at 20,000 ft, then accelerate to Mach 0.85 (533 kt).

Navigation for standard takeoff assumes that the INS has been aligned by gyrocompassing and that GPS is continually available. Navigation for alert takeoff assumes that the INS is aligned by a stored heading reference and that GPS is not available until 5 min after takeoff. In both cases, different turn, climb, and dive maneuvers after takeoff were considered for their effectiveness in reducing alignment errors after takeoff.

The values cited in Table 1 were used to characterize the standard errors (1–σ levels) remaining after gyrocompassing and stored heading alignment respectively. These are the initial conditions for subsequent Kalman filtering of GPS plus inertial sensor outputs.

Three individual test cases are presented here. These are listed in Table 2.

c. Case 1. Case 1 illustrates the potential benefits of using GPS-aided alignment in a standard mission, where the INS is gyrocompassed prior to takeoff, and GPS is available throughout takeoff. A 2/3-*g* turn that produces a 45-deg heading change beginning at $t = 495$ s is included to illustrate the additional improvement available from in-flight maneuvers.

Figure 8 illustrates the history of rms alignment errors. As expected, the availability of GPS during the period of acceleration at takeoff reduces the

Table 1 Alignment errors at takeoff (arcsec)

	Level	Azimuth
Stored heading	40	450
Gyrocompassing	37	225

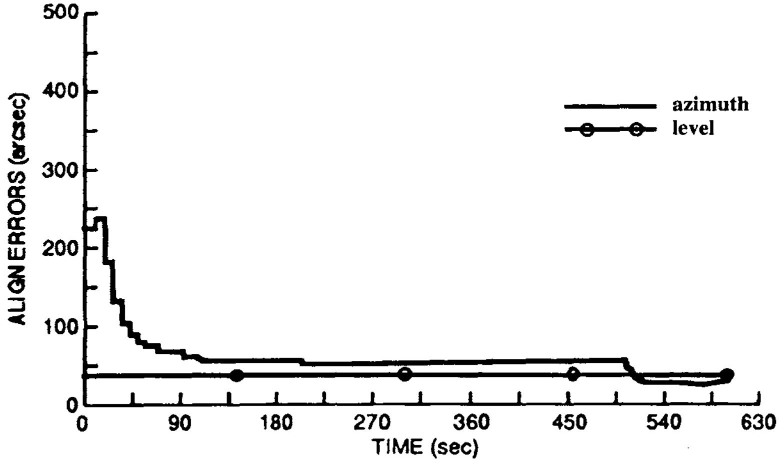


Fig. 8 Inertial navigation systems alignment errors with GPS available at takeoff and gyrocompassed alignment (standard takeoff).

azimuth alignment error; it seems to decrease asymptotically toward about 50 arcsec. The additional acceleration after $t = 495$ s produces a modest further improvement. The final azimuth error is approximately the same as the level error, which signifies that accelerometer bias is the limiting factor in improving alignment by this procedure. An analysis of the error sensitivity of the tilt estimates confirms this intuition, with nonorthogonality of the horizontal accelerometers with respect to the "up" direction being the only other significant error source.

d. Case 2. Case 2 illustrates potential benefits of GPS for a "scramble" takeoff, when GPS is available only after 300 s. The aircraft takes off with a stored heading alignment having a nominal azimuth uncertainty of 450 arcsec. The aircraft executes a 1-g turn that produces a 45-deg heading change starting at $t = 495$ s. (The heading angle for Case 2 is essentially the same as for Case 1, but the turn is completed faster.)

Figure 9 illustrates the history of RMS alignment errors. The error is constant until $t = 300$ s, when GPS becomes available. The availability of GPS makes a

Table 2 Test cases for in-flight alignment studies

Case	Takeoff mode	Alignment mode	GPS on at	Alignment maneuver
1	Standard	Gyrocompass	0 s	2/3-g turn
2	Alert	Stored heading	300 s	1-g turn
3	Alert	Stored heading	300 s	2-g S-turn

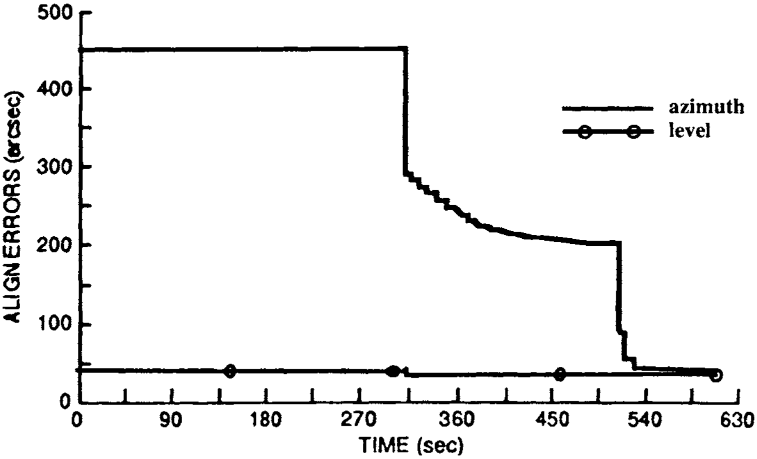


Fig. 9 Inertial navigation systems alignment errors with GPS available 5 min after an alert takeoff.

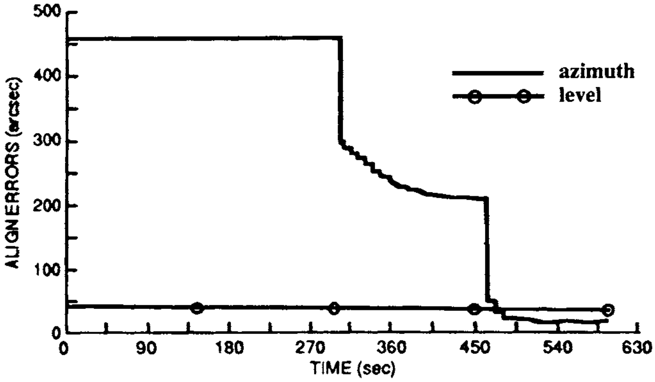


Fig. 10 Inertial navigation systems alignment errors with GPS available at 300 s after an alert takeoff with stored heading alignment.

dramatic improvement in alignment even without maneuvers because the position and velocity errors that have accumulated in the INS navigation results are highly correlated with misalignments. We see that, following an initial decrease to 280 arcsec, the azimuth errors fall to an asymptote of 200 arcsec, which is comparable to gyrocompass alignment accuracy. The execution of a 1-g lateral maneuver drives the azimuth errors to the level predicted by accelerometer bias.

e. Case 3. Case 3 involves an "alert" takeoff with nominal stored heading azimuth errors and a 2-g "s" turn maneuver at t = 460 s. GPS is not available until t = 300 s (corresponding to a five-minute warm-up and acquisition time).

Figure 10 illustrates the rms alignment error. As expected, Figs. 9 and 10 are identical until the maneuver begins, and then the higher g applied in Case 3 reduces the azimuth error below the level for Case 2.

3. *Discussion*

These simulations support the following performance conclusions:

1) In-flight GPS measurements during periods when lateral acceleration is on the order of 1 *g* can be used to reduce azimuthal alignment errors to the same magnitude as level errors. Larger accelerations lead to smaller errors.

2) Even in the absence of lateral acceleration, the availability of GPS measurements can be used to reduce the errors of a stored-heading alignment to the level of gyrocompassing. This reduction exploits the correlation between INS alignment errors and the INS navigation errors that are uncovered by comparison with GPS observations.

3) Simple maneuvers, such as 1-*g* coordinated turns over as little as 45 deg, and lasting for no more than 30 s, are adequate for realizing the benefits of in-flight alignment.

C. Integrated Navigation Solutions During a GPS Outage

The previous examples illustrated certain benefits of GPS/INS integration in a benign environment, with low-to-moderate vehicle dynamics. The use of integration to lengthen the interval that GPS data are available, and to ride out the period of outages caused by the interference or high dynamics is the focus of the following remarks.

1. *Integration Case Study Overview*

In a recent military aircraft avionics upgrade, RCVR 3A was to be integrated with a "navigation-grade" gimbaled-platform INS. The principal performance objective was to limit the error growth after 5 min of any GPS outage, provided that the INS had been calibrated by at least 7 min of GPS measurements prior to the outage. Based on extensive simulations, the following results were determined:

1) The horizontal circular error probable (CEP) after 5 min of GPS outage could be held to 30 m. (Recall that the free-running INS is specified as a 1.0 n.mi./h system.)

2) A 21-state filter mechanized as two independent "horizontal" and "vertical" filters would provide close enough to optimum performance to be a cost-effective, computationally effective approach.

These results were derived from a covariance analysis that incorporated a 73-state truth model for the INS, plus additional error states for GPS.

Table 3 lists these truth states. Table 4 characterizes the mission segments of the standard aircraft flight path used in the simulation.

2. *Navigation Performance of. the Integrated System*

Figure 11a illustrates the growth in rms east-positioning navigation error vs time after the loss of GPS at approximately 2550 s. In this case, the GPS and INS position and velocity measurements were processed by a full optimal 73-state filter with dynamics, plant noise parameters, and measurement noise parameters that exactly match those in the truth model. In other words, this represents the best that any filter can do given this truth model and trajectory. Five minutes after the loss of GPS, the east component of horizontal position errors has grown to 80 ft (24.4-m rms), exclusive of the low-frequency GPS bias. Figure 12 shows the contribution of various error sources to the net navigation error. For at least 10 min, the accelerometer and gravity disturbance terms are dominant. Gyro errors do not become dominant until more than a quarter of a Schuler period (22 min) after the outage.

Figure 11b illustrates the growth in east errors for a suboptimum 21-state filter, whose states are listed in Table 5. These results are almost indistinguishable from Fig. 11a, for the "optimum filter." Finally, Fig. 11c gives the east error for a 15-state horizontal filter whose states are listed in Table 5. These results, too, are nearly indistinguishable from optimum for about the first 420 s of GPS outage.

Table 3 Truth model states

State number	Description	Comment
1,2	Horizontal position error	
3	Wander angle error	
4,5,6	Platform misalignment	100 arcsec initial horizontal 6 deg initial vertical
7,8,9	Velocity error	
10	Vertical position error	
11–13	Auxiliary baro-inertial states	Used in describing baro-inertial loop
14–16	Markov gyro bias	0.002 deg/h rms horizontal 0.005 deg/h rms vertical Correlation time = 5 min
17,18	Markov accelerometer bias	3 μg rms 10 min correlation time
19	Markov baro bias	100 ft rms 10 min correlation time
21–23	Gravity disturbance	35 μg rms 20 n.mi. correlation distance
24–26	Gyro bias	0.01 deg/h rms horizontal 0.022 deg/h rms vertical
27–29	Gyro scale factor error	0.0002 rms
30–32	Gyro misalignments about spin axes	3.3 arcsec rms
33–35	Remaining gyro misalignments	20 arcsec rms
36–41	Gyro *g*-sensitivity	0.015 deg/h/*g* rms
42–44	Gyro *g*-squared-sensitivity	0.02 deg/h/g^2 rms
45–47	Accelerometer bias	150 μg rms
48–50	Accelerometer scale factor error	0.0002 rms
51–53	Accelerometer scale factor asymmetry	0.0001 rms
54–56	Accelerometer nonlinear scale factor asymmetry	10 μg/g^2 rms
57–59	Accelerometer nonlinearity	10 μg/g^2 rms
60–65	Accelerometer orthogonal *g*-squared sensitivity	30 μg/g^2 rms
66–71	Accelerometer misalignments	20 arcsec rms
72	Baroaltimeter bias	300 ft rms
73	Baroaltimeter scale factor error	0.04 rms

Table 4 Test trajectory

Time, s	Description
0–300	Gyrocompass at true heading = 120 deg
300–490	Taxi and turn left approximately 75 deg to prepare for takeoff at a heading of approximately 45 deg
490–540	Takeoff; speed = 434 ft/s, altitude approximately 1000 ft
540–840	Climb; accelerate to 655 ft/s, altitude approximately 9660 ft
840–1020	Level off at approximately 9660 ft; cruise at 655 ft/s
1020–1100	Descend to approximately 200 ft
1100–2720	Level-off at 200 ft; cruise at 655 ft/s
2720–2780	90 deg right turn.
2780–2806	Pop-up, climb to approximately 12,000 ft; accelerate to approximately 820 ft/s
2806–2838	Dive to 635 ft
2838–2858	Level-out at 635 ft; decelerate to 655 ft/s; weapons delivery
2858–2918	90-deg right turn
2918–3226	Low-altitude combat, 300 to 2700 ft; 485 to 655 ft/s
3226–3276	Turning climb to 9900 ft; speed = 655 ft/s
3276–3346	High-altitude combat
3346–5086	Turn toward home; high-altitude cruise at 9900 ft, 655 ft/s
5086–5446	Descend to 3000 ft; accelerate to 820 ft/s
5446–5876	Loiter; decelerate to 434 ft/s
5876–6126	Land

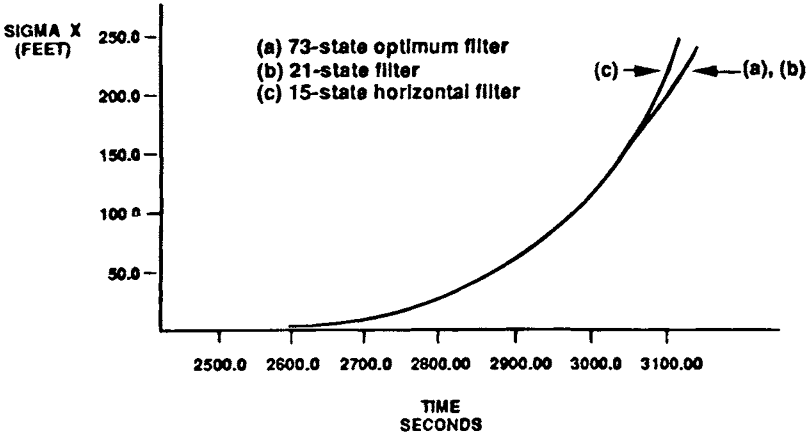


Fig. 11 Root mean square position error vs time. GPS measurements stop at 2550 s.

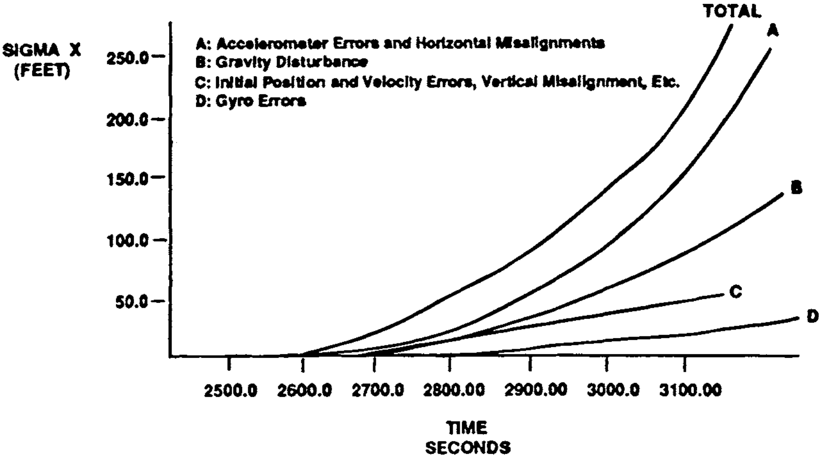


Fig. 12 Contributions to inertial navigation systems error growth during a GPS outage.

Table 5 States of the 15-state horizontal filter

State number	Description	Comment[a]
1,2	Horizontal position error	
3	Wander angle error[a]	
4,5,6	Platform misalignment (4 and 5 also absorb constant accelerometer bias)	Plant noise of 9.4E-17 rad^2/s (horizontal) and 5.9E-16 rad^2/s (vertical) used to account for unmodelled gyro errors
7,8	Horizontal velocity error	Plant noise of 9.E-12 ft^2/s^3 used to account for unmodeled accelerometer error
9–11	Gyro bias	
12,13	Horizontal gravity disturbance (also absorbs other accelerometer errors)	
14,15	Accelerometer scale factor error	

[a] In an actual implementation, wander angle error would be combined into a single state with vertical misalignment.

IV. Summary

We have presented a description of the processes whereby the combination of GPS with an inertial navigator can produce a system performance that is superior to either one acting alone. These benefits include the following:

1) Smaller random errors than seen in stand-alone GPS navigation solutions.

2) Improved availability of GPS operations during maneuvers and in the presence of radio frequency interference (RFI).

3) A navigation solution whose position and velocity errors are bounded by the errors in the GPS navigation solution.

4) A calibrated navigation solution whose errors grow slower than those of a free-running uncalibrated INS during GPS outages.

They are available because the long-term (low-frequency) content of GPS errors is negligible.

All four benefits are available to GPS users who are authorized to obtain the GPS Precise Positioning Service (PPS). Only the first and third are guaranteed to GPS users who are vulnerable to the selective availability clock dither that corrupts the GPS Standard Positioning Service (SPS). Clock dither prevents the tracking loop bandwidths from being reduced to the bandwidth of the INS errors, and corrupts the use of GPS measurements to calibrate the INS error states.

GPS receiver technology is evolving rapidly in response to pressure from the civilian market. This trend is evident in the miniaturization of full-capability receivers that can be physically embedded in a host system, such as an INS or a mission computer. This in turn makes it practical to obtain even higher performance levels by treating both GPS and the INS as sensors that produce measurements to be optimally combined by a navigation filter into an "optimum" navigation solution.

For many users, the primary benefit of embedded configurations is not improved performance. These users focus on savings in volume, weight, power consumption, and cost that are predicted for the embedded system. Some cost savings are nonrecurring (e.g., the one-time investment in developing integration software may be borne by the vendor of the embedded system rather than by the integrator), whereas others are recurring (e.g., reduced production cost of the system hardware).

In addition to these cost and performance benefits, the integrated system may be able to support functional capabilities that were previously not available to the user. For example, the use of in-flight INS alignment can make it possible for an aircraft to take off without waiting for the 5–10 min routinely reserved for INS alignment. This may be a life-saving capability for military aircraft that must take off immediately after an alert that hostile forces are incoming. Another example is the inclusion of error states for gravity anomalies in the integration filter. This is the basis for balloon-borne GPS/INS instrumentation that will allow for more extensive gravity mapping than has ever been possible.[34]

The design of a Kalman integration filter that mechanizes the navigation solution must also address the following standard questions:

1) How many states should be estimated?

2) How often should the filter be updated?

3) How should correlated measurements be treated?

4) By how much can the computational burden of the filter be reduced by exploiting sparseness of the state dynamics matrix or the measurement covariance matrix?

5) How should the filter accommodate transient events such as changes in the constellation of satellites being tracked?

6) How should the filter be tuned to provide robust performance vs unknown aspects of the design problem?

7) How can the filter be made robust against variations in the error characteristics of individual INS or GPS units?

8) How can the filter be used to detect the onset of anomalous conditions that may indicate a failure in the GPS or the INS subsystems?

These design questions are addressed in more detail in the following chapters of this book.

References

[1]Cox, D. B., "Integration of GPS with Inertial Navigation Systems," reprinted in *Collected GPS Papers,* Vol. I, Institute of Navigation, Alexandria, VA, 1980, pp. 144–153.

[2]Johannessen, R., and Asbury, M. J. A., "Towards A Quantitative Assessment of Benefits INS/GPS Integration Can Offer to Civil Aviation," *Navigation,* Vol. 37, No. 4, 1990–1991, pp. 329–346.

[3]Greenspan, R. L., et al., "The GPS Users Integration Guide," *Proceedings of the ION National Technical Meeting* (Santa Barbara, CA), Institute of Navigation, Washington, DC, Jan. 26–29, 1988, pp. 104–112.

[4]Wiederholt, L., and Klein, D., "Phase III GPS Options for Aircraft Platforms," *Navigation,* Vol. 31, No. 2, 1984, pp. 129–151.

[5]Brown, A. K., and Bowles, W. M., "Interferometric Attitude Determination Using GPS," *Proceedings of the Third Geodetic Symposium on Satellite Doppler Positioning* (Las Cruces, NM), Feb. 1982, pp. 1289–1302.

[6]Eller, D., "GPS/IMU Navigation in a High Dynamics Environment," *Proceedings of the First International Symposium on Precise Positioning with GPS* (Rockville, MD), April 15–19, 1985, pp. 773–782.

[7]Bridges, P. D., "Influence of Satellite Geometry Range, Clock and Altimeter Errors, on Two-Satellite GPS Navigation," *Proceedings of the ION GPS '88* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 19–23, 1988, pp. 253–258.

[8]Kalafus, R. M., and Knable, N., "Clock Coasting and Altimeter Error Analysis for GPS," *Navigation,* Vol. 31, No. 4, 1984–85, pp. 289–302.

[9]Bartholomew, R. G., et al., "Software Architecture of the Family of DOD Standard GPS Receivers." *Proceedings of the First Technical Meeting of ION Satellite Division* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 1987, pp. 23–24.

[10]Dellicker, S. H., and Henckel, D., "F-16/GPS Integration Test," *Proceedings of ION GPS-89* (Colorado Spring, CO), Institute of Navigation, Washington, DC, Sept., 1989, pp. 295–303.

[11]Barbour, N., et al., "Inertial Instruments—Where to Now?," *Proceedings of the AIAA Guidance, Navigation, and Control Conference* (Hilton Head, SC), Washington, DC, Aug. 1992 (AIAA Paper 924414-CP).

[12]Boxenhorn, B., et al., "The Micromechanical Inertial Guidance System and Its Application," *Proceedings of the Fourteenth Biennial Guidance Test Symposium* (Holloman, AFB, NM), Oct. 3–5, 1989, Vol. 1, pp. 113–131.

[13]Ward, P., and Rath, J., "Attitude Estimation Using GPS," *Proceedings of the ION National Technical Meeting* (San Mateo, CA), Institute of Navigation, Washington, DC, Jan. 23–26, 1989, pp. 169–178.

[14]Kruczynski, L. R., and Li, P. C., "Using GPS to Determine Vehicle Attitude," *Proceedings of the ION GPS '88* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 19–23, 1988, pp. 139–146.

[15]Jurgens, R., "Real Time GPS Azimuth Determining System," *Proceedings of the ION National Technical Conference* (San Diego, CA), Institute of Navigation, Washington, DC, Jan. 23–25, 1990, pp. 105–110.

[16]van Graas, F., and Braasch, M., "GPS Interferometric Attitude and Heading Determination; Flight Test Results," *Proceedings of the 47th Annual Meeting (ION)* (Williamsburg, VA), Institute of Navigation, Washington, DC, June 10–12, 1991, pp. 183–191.

[17]Graham, W. R., and Johnston, G. R., "Standard Integration Filter State Specification and Accuracy Predictions," *Navigation,* Vol. 33, No. 4, 1986–1987, pp. 295–313.

[18]Cunningham, J., and Lewantowicz, Z. H., "Dynamic Integration of Separate INS/GPS Kalman Filters," *Proceedings of the ION GPS '88* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 19–23, 1988, pp. 273–282.

[19]Diesel, J. W., "Integration of GPS/INS for Maximum Velocity Accuracy," *Navigation,* Vol. 34, No. 3, 1987, pp. 190–211.

[20]Berman, G. J., and Belzer, M., "A Decentralized Square Root Information Filter/Smoother," of the *Proceedings 24th IEEE Conference on Decision and Control* (Ft. Lauderdale, FL), Institute of Electrical and Electronics Engineers, New York, Dec. 1985.

[21]Carlson, N. A., "Federated Square Filter for Decentralized Parallel Processes," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 26, No. 3, 1990, pp. 517–525.

[22]Gelb, A. (ed.), *Applied Optimal Estimation,* MIT Press, Cambridge, MA, 1974, Chap. 4, pp. 102–132.

[23]Maybeck, P. S., *Stochastic Models, Estimation and Control,* Vol. 1, Academic Press, Orlando, FL, 1979.

[24]Bletzacker, F., Eller, D., Forgette, T., Seibert, G., Vavrus, J., and Wade, M., "Kalman Filter Design for Integration of Phase III GPS with an Inertial Navigation System," *Proceedings of the ION National Technical Meeting* (Santa Barbara, CA), Institute of Navigation, Washington, DC, Jan. 26–29, 1988, pp. 113–129.

[25]Lewantowicz, Z. H., and Keen, D. W., "Graceful Degradation of GPS/INS Performance with Fewer than Four Satellites," *Proceedings of the ION National Technical Meeting* (Phoenix, AZ), Institute of Navigation, Washington, DC, Jan. 21–24, 1991, pp. 269–276.

[26]Widnall, W. S., "Stability of Alternate Designs for Rate Aiding of a Noncoherent Mode of a GPS Receiver," Intermetrics Corp., Cambridge, MA, Sept. 25, 1978.

[27]Buechler, D., and Foss, M., "Integration of GPS and Strapdown Inertial Subsystems into a Single Unit," *Navigation,* Vol. 34, No. 2, 1987, pp. 140–159.

[28]Franklin, M., and Pagnucco, S., "Development of Small Embedded GPS/INS RLG and FOG systems," *Proceedings of the ION National Technical Conference* (San Diego, CA), Institute of Navigation, Washington, DC, Jan. 27–29, 1992, pp. 3–12.

[29]Tazartes, D. A., and Mark, J. G., "Integration of GPS Receivers into Existing Inertial Navigation Systems," *Navigation,* Vol. 35, No. 1, 1988, pp. 105–120.

[30] Homer, W. C., "An Introduction to the GPS Guidance Package (GGP)," *Proceedings of the 15th Biennial Guidance Test Symposium, CIGTF Guidance Test Division,* Vol. 1 (Holloman AFB, NM), Sept. 24–26, 1991 pp. 12–15.

[31] Greenspan, R., and Donna, J., "Measurement Errors in GPS Observables," *Navigation,* Vol. 33, No. 4, 1986–1987, pp. 319–334.

[32] Soltz, J. A., Donna, J. I., and Greenspan, R. L., "An Option for Mechanizing Integrated GPS/INS Solutions," *Navigation,* Vol. 35, No. 4, 1988–1989, pp. 443–458.

[33] "Specification for USAF Standard Form, Fit and Function (F^3) Medium Accuracy Inertial Navigation Unit," SNU-84-1, Rev. 3, Amendment, Feb. 13, 1987, USAF Aeronautical Systems Division (now Aeronautical Systems Center).

[34] Jekeli, C., Doyle, T., Nicolaides, P., and Galdos, J., "Instrumentation Design and Analysis for Balloon-Borne Gravimetry and Attitude Determination Using GPS and INS," *Proceedings of the ION National Technical Meeting* (San Francisco, CA), Institute of Navigation, Washington, DC, Jan. 20–22, 1993, pp. 519–533.

Chapter 8

Receiver Autonomous Integrity Monitoring Availability for GPS Augmented with Barometric Altimeter Aiding and Clock Coasting

Young C. Lee*
MITRE Corporation, McLean, Virginia 22102

I. Introduction

ONE of the most important criteria for the operational approval of the use of Global Positioning System (GPS) for civil air navigation for instrument flight rules (IFR) operations is safety, which requires assuring the integrity of navigation solutions derived from GPS. Integrity monitoring detects erroneous information not detectable by other means, such as signal disappearances or failures identified through the satellite health bit flag. An integrity monitoring method for GPS that does not require an external monitoring system, and thus, is simple and practical to implement is receiver autonomous integrity monitoring (RAIM). RAIM, which is discussed in another chapter in this book, is to perform detection (detection of position error beyond protection limits) and identification (identification of failed satellite) functions.[1] For near term use of GPS as a supplemental system in the National Airspace System (NAS), the RAIM detection function is essential; on the other hand, the RAIM identification function, although not essential, is highly desirable. Without the identification function, a satellite failure would cause an outage of the position fixing function. High availability of the navigation and RAIM detection functions is important because when this function cannot be performed, the user must fall back on another approved navigation system. High availability of the navigation and RAIM functions is even more important for GPS to be authorized to provide required navigation performance (RNP), because a failure of a GPS-based system would have an impact over a large area, and thus, have a serious operational impact because of

*Lead Engineer, Center for Advanced Aviation System Development, MS W307, 7525 Colshire Drive.

the lack of a backup system. (An RNP system is one that satisfies all requirements for flight in a given airspace.)

As a means of improving availability of the navigation and RAIM functions, GPS augmentations in the form of barometric altimeter aiding and clock coasting are often suggested. Implementation of these augmentations would be relatively simple and could be done in a timely manner. It is noteworthy that a requirement for barometric altimeter aiding was included in the FAA's Technical Standard Order (TSO)-C129,[2] which specifies requirements for GPS avionics equipment to be used as a supplemental means of navigation for en route (domestic and oceanic), terminal, and nonprecision approach phases of flight. This paper presents the technical analyses of availability and outage duration statistics of navigation and RAIM detection and identification functions for GPS with the altimeter aiding and/or clock coasting for en route, terminal, and nonprecision approach phases of flight.

The approach used in this paper is to compute temporal characteristics of both navigation and RAIM availabilities over three major domestic and oceanic routes. In addition, navigation and RAIM availabilities for users making a final approach was evaluated for five major airport locations in the conterminous United States (CONUS). The constellations considered include the Optimal 21 constellation (Opt21), 24 Primary constellation (24Pr), and the 24 Primary constellation with a typical set of 3 satellites failed (24Minus3). Because the Department of Defense (DOD) guarantees at least 21 operating satellites 98% of the time, but 24 operating satellites only 70%, it was considered important to determine availability for different constellations with less than 24 operating satellites. In particular, the 24Minus3 constellation is considered to give a reasonable lower bound in terms of the expected RAIM availability under published DOD policy.

Approaches for the augmentation are described first. This is followed by the analytical results. Finally, the findings of the analyses are summarized. The annex, which is the contribution of Rolf Johannessen and Charles Dixon, describes statistical distribution of the barometric error growth as a function of distance flown. This error growth rate has a significant impact on the effectiveness of barometric altimeter aiding, as discussed later.

II. Methods of Augmentations

This section describes a procedure for augmenting the basic GPS solution equation; first for the case of barometric altimeter aiding, and then for the case of clock coasting. One basic assumption in the following discussion is that GPS will be augmented with the baroaltitude or clock only when necessary; that is, if navigation or RAIM functions can be provided with GPS alone at a particular point in time, no augmentation is to be incorporated.

A. Augmented Geometry for Barometric Altimeter Aiding

In this method, barometric altimeter data provide an additional piece of information for navigation, RAIM detection, and RAIM identification functions. The barometric altimeter data are incorporated into the basic GPS position solution.

The basic equation for the set of GPS measurements, without any augmentation, is a linearized equation of the form

$$\begin{bmatrix} \Delta\rho_1 \\ \Delta\rho_2 \\ \cdots \\ \Delta\rho_n \end{bmatrix} = G \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta t \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix} \tag{1}$$

where $\Delta\rho_i$ is the difference between the measured pseudorange and predicted pseudorange based on the nominal GPS position and clock bias; G is $n \times 4$ geometrical connection matrix (directional cosines of line-of-sights to satellites) relating the measurements to the three components of position, and time; Δx, Δy, Δz, and Δt are perturbations in x, y, and z coordinates and time t from the nominal position and clock bias; and ϵ_i is range measurement error for the ith satellite.

In the case of barometric altimeter aiding, the above equation is augmented by a linear equation of the following form, making use of the baroaltitude measurement.

$$\Delta B = [0\ 0\ 1\ 0] \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta t \end{bmatrix} + \epsilon_{\text{baro}} \tag{2}$$

where ΔB is the difference between the measured barometric altitude (calibrated with either GPS as described below or the local barometric setting) converted to WGS-84 altitude, and the predicted altitude based on the nominal GPS position; and ϵ_{baro} is the error in measured barometric altitude after it has been calibrated (either by GPS or local pressure setting) and converted to WGS-84 altitude.

In a conventional RAIM algorithm such as the least-squares residual RAIM method, it is assumed that all independent sensor measurement errors have the same variance. Therefore, in case the standard deviation of ϵ_{baro} is different from σ_{sv}, the above equation should be properly scaled before it is added to the basic GPS measurement equation. That is,

$$\frac{\Delta B}{(\sigma_{\text{baro}}/\sigma_{sv})} = \begin{bmatrix} 0 & 0 & \dfrac{1}{(\sigma_{\text{baro}}/\sigma_{sv})} & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta t \end{bmatrix} + \frac{\epsilon_{\text{baro}}}{(\sigma_{\text{baro}}/\sigma_{sv})} \tag{3}$$

where σ_{baro} is the standard deviation of ϵ_{baro}. Function availability is determined from the expanded coefficient matrix consisting of matrix G and the coefficient matrix in Eq. (3).

Two different ways of aiding GPS with barometric altimeter have been analyzed because of their utility for aviation applications. One way is to use pressure altitude data (altitude measurements with no correction for local temperature or pressure), but calibrated with GPS. The other way is to use altitude data calibrated with a local barometric setting.

B. Barometric Altimeter Aiding with GPS-Calibrated Pressure Altitude Data

This case uses altimeter data with a standard setting and no correction for temperature or local pressure, using these steps.

1) When navigation and RAIM detection (or identification) functions can be provided with GPS alone, baroaltitude is not used for these functions. At these times, if the geometry is good so that integrity can be assured to a certain level for the vertical position estimate from GPS, the barometric altimeter data are calibrated; that is, the offset between the altitude from the altimeter and altitude estimated from GPS is recorded. The geometric criterion that defines goodness for barometric altimeter data calibration is discussed below in conjunction with barometric altimeter calibration error.

2) If navigation and RAIM functions are not available with GPS alone, GPS is augmented by the measured barometric altimeter data calibrated with the offset obtained in step 1 at the most recent time at which geometry met the criterion for calibration. The augmented geometry for barometric altimeter aiding is used with the proper value of σ_{baro}.

In step 2, whatever value we use for σ_{baro} makes an impact on the availability analysis results. Most RAIM analyses to date for barometric altimeter aiding[3–5] have assumed a certain fixed value of σ_{baro} considered typical for each phase of flight. To model σ_{baro} more realistically, a different approach is taken in the present analysis. That is, the two sources of the error in the barometric altimeter data are separately evaluated before they are combined for the calculation of σ_{baro}: calibration error, and error caused by a change of the offset between the pressure altitude and geometric altitude. Each of these error sources is discussed below. Because these are independent sources of error, σ_{baro} can be obtained from their rss.

1. *Barometric Altimeter Calibration Error Caused by GPS*

This is the error associated with calibration of barometric altimeter data with GPS. There are two questions involved in this GPS calibration of barometric altimeter data. First, under what conditions can we calibrate barometric altitude? Second, how is the calibration error to be estimated?

To answer the first question, we can calibrate barometric altitude when the RAIM integrity algorithm ensures integrity (i.e., no alarm). If none of the satellites is failing with an abnormally large range error, the following should hold:

$$\sigma_h = (\text{all-in-view-set VDOP}) \times (1\ \sigma\ \text{pseudorange error}) \tag{4}$$

where σ_h is the standard deviation of the GPS vertical position error. In this case, we can take σ_h as the standard deviation of barometric altimeter calibration error.

It should be noted, however, that before calibrating barometric altimeter data with GPS vertical position data, the validity of the GPS vertical data must be assured. Of course, the calibration will be done only when RAIM does not raise an alarm. However, we should remember that no alarm in RAIM does not mean every satellite has a nominal value of error; it only means that none of the satellite range errors has become large enough to cause the horizontal position error to violate the protection limit for the given phase of flight. Therefore, it is possible that one of the satellites is beginning to fail with a large error and has not yet been detected by the RAIM algorithm. In this case, barometric altimeter data would be calibrated with erroneous GPS vertical data. When this erroneous barometric altimeter data are subsequently used to augment GPS, the RAIM capability cannot be trusted. For this reason, a proper amount of calibration error should be assigned that would reflect the level of integrity (i.e., protection limit) in the GPS vertical position data. The following is the approach used to account for calibration error.

Earlier analyses of RAIM performance show that the level of integrity (i.e., protection limit) that can be assured is strongly correlated with the worst subset horizontal dilution of precision (HDOP) and position dilution of precision (PDOP) (i.e., the maximum HDOP or PDOP among $n - 1$ out of the n satellites used for navigation). Specifically, it is shown in Refs. 6 and 7 that if the worst subset HDOP is less than or equal to three, then a horizontal error of greater than approximately 300 m can be protected against. This is under the assumption of the presence of selective availability (SA) with a standard deviation of 33 m for the pseudorange error. From this, it is observed that

$$\text{Horizontal protection limit} = 3 \times (\text{worst subset HDOP}) \times (1\ \sigma\ \text{pseudorange error}) \tag{5}$$

By extrapolating the above to the case of vertical position integrity, we obtain

$$\text{Vertical protection limit} = 3 \times (\text{Worst subset VDOP}) \times (1\ \sigma\ \text{pseudorange error}) \tag{6}$$

The Radio Technical Commission for Aeronautics (RTCA) specification[8] requires a missed detection probability of 0.001. This means that when there is a degradation of satellite ranging data, the probability that the protection limit is exceeded will be less than 0.001. Although it is difficult to characterize the statistical distribution of the error in the vertical position estimate used for calibration, we can assume that it has a normal distribution with a standard deviation of the error determined from the missed detection probability of 0.001. Under this assumption, the protection limit would correspond to 3.3 times the standard deviation of the error, which is approximated as 3 instead of 3.3. Therefore,

$$\sigma_{\text{baro calib}} = \text{worst subset vertical dilution of precision (VDOP)} \times (1\ \sigma\ \text{pseudorange error}) \tag{7}$$

One thing to note is that although the barometric altimeter data should be calibrated whenever possible, it is better not to calibrate them when $\sigma_{\text{baro calib}}$ is too large. This is because the barometric altimeter calibration error may increase faster than the error caused by the pressure variation as a result of level or nonlevel flight. For this reason, it is assumed in the current analysis that barometric altimeter data are calibrated only when the worst subset VDOP is smaller than 5, which corresponds to a vertical protection limit of 500 m with SA on.

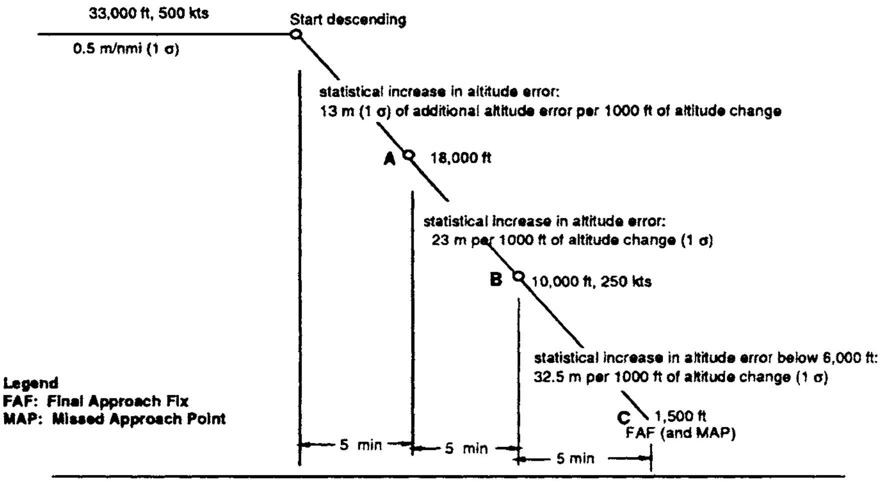
As a further measure to ensure that pressure altitude data are not calibrated by contaminated GPS data, TSO-C129 requires that calibration of pressure altitude data by GPS be performed only when the GPS test statistic is less than a threshold that corresponds to the 95th percentile point of the distribution of the test statistic in the presence of SA and that no other errors are present. This measure, along with the constraint on the maximum subset VDOP, would guarantee a probability of less than 0.001 of calibration with contaminated GPS data.

2. *Error Caused by Change of Offset Between Pressure Altitude and Geometric Altitude*

As the aircraft flies along after the barometric altimeter is calibrated with GPS, the offset between the altitude given by the barometric altimeter and the geometric altitude varies with changes in local pressure and temperature, as well as due to any change in aircraft altitude.[9,10] The appendix to this chapter describes statistical distribution of altitude offset vs distance flown. The growth of this uncertainty in barometric altimeter data varies widely, depending on the flight profile. In the current analysis, two different types of flight are considered, level flight and descending flight.

a. Level flight. Table A.1 in the appendix indicates that the standard deviation of the growth rate is approximately 0.5 m and 0.2 m per n.mi. of level flight at 250 and 850 mbars, respectively. A similar analysis in Ref. 10 determines that the worst growth rate of the uncertainty is 1 m per n.mi. of level flight. On the basis of these two sources, the standard deviation of the growth rate is estimated somewhat conservatively at 0.5 m per n.mi.; this value is used in the current analysis irrespective of the altitude. Although using the larger value of the rate may result in somewhat lower availability of functions, it should not affect integrity performance negatively.

b. Descending flight. For this case, a flight profile illustrated in Fig. 1 is assumed. The rates of growth of uncertainty shown in the figure are based on the results of an analysis reported in Ref. 10. (The rates shown may be considered


33,000 ft, 500 kts
0.5 m/nmi (1 σ)
Start descending
statistical increase in altitude error:
13 m (1 σ) of additional altitude error per 1000 ft of altitude change
A
18,000 ft
statistical increase in altitude error:
23 m per 1000 ft of altitude change (1 σ)
B
10,000 ft, 250 kts
statistical increase in altitude error below 6,000 ft:
32.5 m per 1000 ft of altitude change (1 σ)
C
1,500 ft
FAF (and MAP)
Legend
FAF: Final Approach Fix
MAP: Missed Approach Point
5 min
5 min
5 min

Notes:
1. Local pressure input is assumed to be used starting at 10,000 ft.
2. Three points of RAIM availability evaluation: A, B, and C

Fig. 1 Flight profile model.

representative for pressure altitude with no temperature correction. It might be possible to reduce the rates significantly with temperature correction of pressure altitude.)

3. *Procedure to Calculate Function Availability*

Figure 2 outlines the procedure to calculate function availability for barometric altimeter aiding. It is shown that the barometric altimeter is calibrated only if the detection function is available and the worst subset VDOP is less than a certain threshold.

C. Barometric Altimeter Aiding with Local Pressure Input

This is the case in which the barometric altimeter data with local barometric setting are used with no GPS calibration process. For σ_{baro}, the following values are used, based on Ref. 10: $\sigma_{\text{baro}} = 290$ m at 10,000 ft; and $\sigma_{\text{baro}} = 49$ m at a typical assumed minimum descent altitude of 1500 ft.

As defined earlier, σ_{baro} is the standard deviation of ϵ_{baro}, which is the error in measured barometric altitude after it has been calibrated (either using GPS altitude or local pressure setting) and converted to WGS-84 altitude. It is noted that there exists a finite amount of difference (e.g., maximum of about 50 m over CONUS) between the WGS-84 ellipsoid altitude and mean sea level altitude as a function of location. TSO-C129 (Ref. 2) requires that this difference be taken into account below 18,000 ft geometric altitude in calibrating the pressure altitude corrected for the local barometric pressure setting. It is believed that with this provision in TSO-C129, the above σ_{baro} values, especially the first, are conservative.

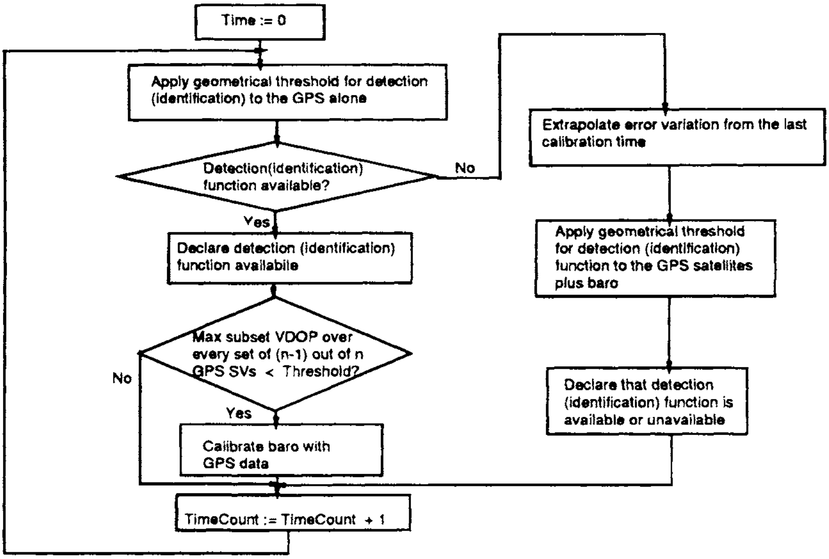

Time := 0
Apply geometrical threshold for detection (identification) to the GPS alone
Detection(identification) function available?
No
Yes
Declare detection (identification) function availabile
Max subset VDOP over every set of (n-1) out of n GPS SVs < Threshold?
No
Yes
Calibrate baro with GPS data
TimeCount := TimeCount + 1
Extrapolate error variation from the last calibration time
Apply geometrical threshold for detection (identification) function to the GPS satellites plus baro
Declare that detection (identification) function is available or unavailable

Fig. 2 Flow chart used to determine availabilities of detection and identification functions for the case of baro aiding with GPS calibration.

D. Augmented Geometry for Clock Coasting

The procedure for clock coasting is similar to that for barometric altimeter aiding with pressure altitude. That is, the user receiver clock is calibrated by GPS under good geometry defined below. When GPS satellite geometry is no longer good enough to provide reliable RAIM functions, GPS is augmented with a clock time measurement. The user receiver clock time is incorporated into the solution for navigation and RAIM functions as outlined below.

An additional equation to be included in Eq. (1) from the user receiver clock time measurement is expressed as:

$$\Delta T = [0 \quad 0 \quad 0 \quad 1] \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta t \end{bmatrix} + \epsilon_{\text{clock}} \tag{8}$$

where ΔT is the difference between the measured user receiver clock time calibrated with GPS and the nominal GPS clock time, and ϵ_{clock} is the error in user receiver clock time (calibrated with GPS). If the standard deviation of of ϵ_{clock} is not equal to σ_{sv}, the above equation should be scaled before it is added to Eq. (1):

$$\frac{\Delta T}{(\sigma_{\text{clock}}/\sigma_{sv})} = \left[0 \quad 0 \quad 0 \quad \frac{1}{(\sigma_{\text{clock}}/\sigma_{sv})} \right] \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \\ \Delta t \end{bmatrix} + \frac{\epsilon_{\text{clock}}}{(\sigma_{\text{clock}}/\sigma_{sv})} \tag{9}$$

where σ_{clock} is the standard deviation of ϵ_{clock}; σ_{clock} is determined by calibration error and error caused by receiver clock drift. Because these are independent sources of error, σ_{clock} can be obtained from their rss. Each of these error sources is discussed below.

1. Clock Calibration Error

As in the case for the barometric altimeter calibration error, calibration of the receiver clock requires integrity in the GPS time estimate. The level of integrity that can be assured in this time estimate is strongly correlated with the worst time dilution of precision (TDOP) among $(n - 1)$ out of n satellites used for navigation. Following logic similar to that for the barometric altimeter calibration,

$$\sigma_{\text{clock calib}} = (\text{worst subset TDOP}) \times (1\ \sigma \text{ pseudorange error}) \tag{10}$$

As in the case of barometric altimeter aiding, if calibration cannot be done with a certain tight integrity level, it is better not to calibrate the receiver clock. In the current analysis, the receiver clock was calibrated by GPS time only when the worst subset TDOP was less than or equal to three.

2. Error Because of Receiver Clock Drift

Uncertainty in the receiver clock time grows because of clock drift from GPS calibration time until clock aiding time. It is assumed that the growth rate is independent of the flight profile. A drift of 10^{-9} was assumed, which corresponds to a good temperature-controlled crystal oscillator. It is assumed here that there is no attempt to estimate the clock drift rate. Although estimation of the clock rate may potentially be promising, the presence of SA would limit the accuracy with which the clock rate can be estimated.

3. Procedure to Calculate Function Availability

The procedure to calculate function availability for clock coasting is similar to the procedure for barometric altimeter aiding shown in Fig. 2.

E. Simultaneous Use of Barometric Altimeter Aiding and Clock Coasting

If barometric altimeter aiding and clock coasting are used simultaneously, both equations of the form in Eqs. (3) and (9) are included in the augmented solution. The same set of model parameters and thresholds are used as that for the individual cases of barometric altimeter aiding and clock coasting.

III. Definitions of Function Availability

Availability is defined as the percentage of time that the system provides the required performance of the function for the phase of flight. In this analysis, availability is calculated on the basis of whether the user-to-satellite geometry (or an augmented geometry) is good enough to provide a satisfactory level of performance. The performance requirements for each of the functions are described below.

A. Navigation Function

Availability of the navigation function is based on the 2 drms (twice the distance root mean square) horizontal position radial error value calculated from the well-known equation

$$2 \text{ drms} = 2 \times \text{HDOP} \times \sigma_{sv} \tag{11}$$

where σ_{sv} is the standard deviation of satellite range error and HDOP is horizontal dilution of precision. That is, if the geometry gives a small enough HDOP for the all-in-view satellite solution to make 2 drms less than a given position error tolerance, then the navigation function is considered available for that geometry.

B. Receiver Autonomous Integrity Monitoring Detection Function

Availability of the RAIM detection function is based on the requirements adopted by RTCA SC-159 for the supplemental use of GPS[8]: presence of a malfunction of a satellite causing the position error protection limit to be violated shall be detected with a minimum probability of 0.999 given that the protection limit is violated; and the rate of alarms (false or true) will not be more than 0.002/h. An empirical relationship that determines, for any given geometry, availability/unavailability of RAIM detection function performing within the level specified above is established via extensive Monte Carlo simulations in

Ref. 11. This relationship, which is based on what is called approximate radial-error-protected (ARP), is used in the current availability analysis.

C. Receiver Autonomous Integrity Monitoring Identification Function

No formal requirements for the RAIM identification function have been established. For this reason, the following assumption is made. Upon occurrence of a malfunctioning satellite with an abnormal range error, RAIM will be able to detect the occurrence and also to correctly identify the satellite before the protection limit is violated. This detection and identification function must be accomplished with a probability of 0.999.

The geometric criteria that satisfy the above requirement are derived from a study reported in Ref. 12, which developed a new identification algorithm and showed, via simulation, that the above identification function performance requirement can be met if PDOP is smaller than a PDOP threshold for every set of $n - 2$ out of n satellites in view where

$$\text{PDOP threshold} = 25 \times [\text{protection limit (n.mi.)}] \times (33/\sigma_{sv}) \qquad (12)$$

This is the criterion used in the current analysis.

IV. Results

This section first describes the cases analyzed and parameters evaluated. Then, the results are presented and discussed. Many different cases were analyzed in terms of such factors as constellation, protection limits, and user locations. These are summarized in Table 1. As noted in the table, the user-to-satellite geometries have been sampled every 5 min.

A. Parameters of Interest

1. Availability

As stated earlier, availability is defined as the percentage of time that the system provides the required performance for the phase of flight. For the case of a user on one of the routes, availability is the average over the duration of flight and over six departure times. For the case of a user at an airport area, the availability is the average over a 24-h period.

2. Average Duration of Outage

This is the average duration of outages of the respective function (navigation, RAIM detection, or RAIM identification) over a flight or over 24-h period.

3. Maximum Duration of Outage

For an airport area, this is the maximum outage duration observed over a 24-h period. For flight along a route, this maximum duration of outage is the maximum of the maximum outage duration in each flight over flights with six different departure times.

Table 1 Summary of analysis scenarios

Constellation	User locations
Optimal 21[a]	User on a moving platform[d]
24 Primary[a]	New York (JFK)–Los Angeles (LAX)
24 Primary with 3 satellites failed (satellites 1, 10, and 11)[b]	San Francisco (SFO)–Narita, Japan (NRT)
	Dallas (DFW)–Paris (CDG)
Protection limits[c]	Airport locations[e]
0.3 n.mi.	San Francisco (SFO)
1 n.mi.	Dallas-Fort Worth (DFW)
2 n.mi.	Chicago (ORD)
12.6 n.mi.	New York (JFK)
Standard deviation	Atlanta (ATL)
of pseudorange error	
σ = 33 m (SA present)	
σ = 10 m (SA absent)	

[a] 21 and 24 satellite positions defined in Ref. 14 but with repeating ground tracks.

[b] This is considered to be an "average" 21 satellite constellation in terms of coverage.[15]

[c] The first three protection limits are RTCA requirements[8] for the nonprecision approach, terminal, and en route phases of flight, respectively. The last protection limit of 12.6 n.mi. corresponds to the current requirement for the oceanic phase of flight.

[d] It is assumed that the user aircraft is flying at a speed of 500 n.mi./h over the great circle defined by each pair of locations. User-satellite geometries are sampled every 5 min. Availability is obtained as an average over six departure times, spaced four hours apart.

[e] These airport locations in CONUS were used to evaluate availability, for terminal and nonprecision approach phases of flight. Availability is obtained as an average over 24-h period with user-to-satellite geometries sampled every 5 min.

B. Discussion of Results

Referring to Tables 2–4, availabilities of navigation, detection, and identification functions, respectively, for protection limits of 1, 2, and 12.6 n.mi. are shown for the three cases of no augmentation, barometric altimeter aiding with GPS calibration, and clock coasting (The combined use of barometric altimeter aiding and clock coasting is treated later in Tables 6 and 7.) For barometric altimeter aiding, level flight was assumed, which is typical for the oceanic and the majority of the en route phases of flight. The results show the following:

1) For navigation, all three constellations have comparable very high availabilities. In fact, only 24Minus3 without an altimeter or clock augmentation suffers any outage at all and has average and maximum outage durations up to 15 and 25 min, respectively.

2) For detection, 24Pr and Opt21 have comparable availability, and 24Minus3 has significantly degraded availability. For the 24Minus3 constellation, average and maximum outage durations are mostly 15–40 min and 50–80 min, respectively. For the other two constellations, they are typically 10 and 25 min, respectively.

3) For identification, only augmented 24Pr has an availability that might be acceptable for a sole-means of navigation (i.e., without other augmentations, such as inertial systems). For the 24 Pr constellation, average and maximum outage durations are typically 5–10 min and 15–25 min, respectively. For the

Table 2 Availabilities of navigation function, SA on, mask angle = 7.5 deg

	24 Primary			Optimal 21			24 Pr minus 3		
	A	B	C	A	B	C	A	B	C
	Protection limit of 1 n.mi.								
GPS alone									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	99.1	8	15
DFW–CDG	100	0	0	100	0	0	99.2	25	25
Baro aiding with GPS calibration in level flight									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	99.8	5	5
DFW–CDG	100	0	0	100	0	0	100	0	0
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	99.7	5	5
DFW–CDG	100	0	0	100	0	0	100	0	0
	Protection limit of 2 n.mi.								
GPS alone									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	99.1	8	15
DFW–CDG	100	0	0	100	0	0	99.2	25	25
Baro aiding with GPS calibration in level flight									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	100	0	0
DFW–CDG	100	0	0	100	0	0	100	0	0
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	99.8	5	5
DFW–CDG	100	0	0	100	0	0	100	0	0
	Protection limit of 12.6 n.mi.								
GPS alone									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	99.4	10	15
DFW–CDG	100	0	0	100	0	0	99.2	25	25
Baro aiding with GPS calibration in level flight									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	100	0	0
DFW–CDG	100	0	0	100	0	0	100	0	0
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	100	0	0	100	0	0
SFO–NRT	100	0	0	100	0	0	100	0	0
DFW–CDG	100	0	0	100	0	0	100	0	0

A = Availability, %.
B = Average outage duration, min.
C = Maximum outage duration, min.

Table 3 Availabilities of receiver autonomous integrity monitoring detection function, SA on, mask angle = 7.5 deg

	24 Primary			Optimal 21			24 Pr minus 3		
	A	B	C	A	B	C	A	B	C
	Protection limit of 1 n.mi.								
GPS alone									
JFK–LAX	100	0	0	99.7	5	5	88.5	14	80
SFO–NRT	99.4	5	5	94.1	10	25	91	17	85
DFW–CDG	99.5	8	10	94.1	8	15	92.9	12	55
Baro aiding with GPS calibration in level flight									
JFK–LAX	100	0	0	100	0	0	93.9	48	80
SFO–NRT	99.8	5	5	100	0	0	94.3	31	85
DFW–CDG	99.8	5	5	99.7	10	10	96.3	16	50
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	100	0	0	92.6	23	80
SFO–NRT	99.8	5	5	97.7	9	20	92.7	21	85
DFW–CDG	99.7	5	5	97.4	11	15	94.2	12	55
	Protection limit of 2 n.mi.								
GPS alone									
JFK–LAX	100	0	0	100	0	0	91.3	17	80
SFO–NRT	99.8	5	5	96.9	7	15	93.7	16	80
DFW–CDG	99.8	5	5	96.8	6	10	95	12	55
Baro aiding with GPS calibration in level flight									
JFK–LAX	100	0	0	100	0	0	94.9	40	75
SFO–NRT	99.8	5	5	100	0	0	95.7	20	50
DFW–CDG	100	0	0	100	0	0	98.1	15	35
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	100	0	0	93.6	33	80
SFO–NRT	99.8	5	5	99.8	5	5	94.6	22	80
DFW–CDG	100	0	0	99	6	10	96.5	16	55
	Protection limit of 12.6 n.mi.								
GPS alone									
JFK–LAX	100	0	0	100	0	0	92.9	22	80
SFO–NRT	100	0	0	99.4	7	10	94.9	24	80
DFW–CDG	100	0	0	99.2	5	5	97	16	50
Baro aiding with GPS calibration in level flight									
JFK–LAX	100	0	0	100	0	0	96.8	50	50
SFO–NRT	100	0	0	100	0	0	98.8	8	15
DFW–CDG	100	0	0	100	0	0	99.2	25	25
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	100	0	0	96.8	25	35
SFO–NRT	100	0	0	100	0	0	98.6	23	30
DFW–CDG	100	0	0	100	0	0	98.4	25	40

A = Availability, %.
B = Average outage duration, min.
C = Maximum outage duration, min.

Table 4 Availabilities of receiver autonomous integrity monitoring identification function, SA on, mask angle = 7.5 deg

	24 Primary			Optimal 21			24 Pr minus 3		
	A	B	C	A	B	C	A	B	C
	Protection limit of 1 n.mi.								
GPS alone									
JFK–LAX	87.8	10	15	51	19	65	57.7	23	125
SFO–NRT	85	12	35	55.7	17	85	53.4	27	200
DFW–CDG	89.9	9	25	56.9	18	110	62.2	24	180
Baro aiding with GPS calibration in level flight									
JFK–LAX	99.4	5	5	94.2	10	20	80.4	24	100
SFO–NRT	98.9	6	10	90.9	11	35	83.3	20	145
DFW–CDG	99.2	6	10	92	8	20	88.1	17	125
Clock coasting (1E-9 drift)									
JFK–LAX	98.7	7	10	92	9	25	76.9	21	100
SFO–NRT	97.1	12	25	85.6	13	65	76.5	25	185
DFW–CDG	98.9	9	15	86.9	12	60	82.9	21	180
	Protection limit of 2 n.mi.								
GPS alone									
JFK–LAX	93.9	7	10	64.4	12	60	67	17	110
SFO–NRT	91.2	7	30	71.1	11	60	64.8	16	190
DFW–CDG	94.4	7	15	68.6	12	70	71.8	14	140
Baro aiding with GPS calibration in level flight									
JFK–LAX	99.7	5	5	97.1	9	15	86.5	19	95
SFO–NRT	99.1	6	10	95.7	9	20	87.2	18	125
DFW–CDG	99.4	5	5	96.3	6	15	91	14	75
Clock coasting (1E-9 drift)									
JFK–LAX	99.7	5	5	96.5	7	15	84	21	95
SFO–NRT	98.6	6	15	93.7	14	50	82.7	26	175
DFW–CDG	99.4	5	5	91.5	10	40	87.7	16	125
	Protection limit of 12.6 n.mi.								
GPS alone									
JFK–LAX	99	5	5	90.1	7	20	77.6	19	105
SFO–NRT	97.4	9	25	85.3	11	50	78.2	19	155
DFW–CDG	98.4	7	15	86.5	15	60	83.5	16	140
Baro aiding with GPS calibration in level flight									
JFK–LAX	99.7	5	5	99	5	5	92.3	17	80
SFO–NRT	99.7	5	5	98.9	6	10	93.5	15	85
DFW–CDG	100	0	0	99.2	5	5	96	10	50
Clock coasting (1E-9 drift)									
JFK–LAX	100	0	0	99.7	5	5	92.3	24	80
SFO–NRT	99.8	5	5	98.5	6	10	94	20	100
DFW–CDG	99.8	5	5	98.9	6	10	95	10	50

A = Availability, %.
B = Average outage duration, min.
C = Maximum outage duration, min.

other two constellations, outage durations are much longer; for example, maximum outage duration for 24Minus3 is typically 100–200 min.

It should be noted that although barometric altimeter aiding always improves availability, the improvement is most dramatic for the identification function, when availability is fairly low without barometric altimeter aiding (e.g., for identification function for a protection limit of 1 n.mi.). On the other hand, compared with GPS-calibrated barometric altimeter aiding in level flight, clock coasting does not improve availabilities as much, even with the assumption of a clock drift corresponding to a high-quality, temperature-controlled crystal oscillator. It should be noted, however, that a program is currently underway to develop an inexpensive microminiature atomic clock with a drift of 10^{-11} or less within the next five years.[13] If such an accurate clock is incorporated into a GPS receiver for aviation users, clock coasting would increase RAIM availability significantly because for a given error, the coasting time increases by a factor of 100 relative to that of a crystal oscillator. Also, RAIM performance and RAIM availability could be improved if the clock drift rate is also estimated along with the clock bias. In that case, however, the presence of SA would limit the accuracy with which the clock rate can be estimated.

For the 24Minus3 constellation, Table 5 compares availabilities at five major airport areas for three different augmentation cases: 1) GPS-calibrated barometric altimeter aiding in descending flight; 2) barometric altimeter aiding using local pressure input; and 3) clock coasting. It is observed that in descending flight at 10,000 ft altitude, barometric altimeter aiding using local pressure input brings more improvement in RAIM availability than the case of GPS-calibrated barometric altimeter aiding. This difference in improvement is much more pronounced at 1500 ft altitude. The table also shows that in descending flight, the availability improvement with GPS-calibrated barometric altimeter aiding is comparable to the case of clock coasting with a receiver clock having a drift assumed in the current analysis. For both cases, however, the results for RAIM identification show very low availabilities.

In Table 6, the availabilities for the case of SA off is compared with those for the case with SA on for the 24Minus3 constellations. As expected, turning SA off always brings some improvement in availability. According to results not reported in this paper, the improvement is much more significant for a protection limit of 0.3 n.mi. Also shown in the table are the case of simultaneous use of barometric altimeter aiding and clock coasting. It is observed that using both barometric altimeter aiding and clock coasting always improves the availability to at least the better of the individual availabilities and often somewhat more than that.

In Table 7, for the 24Minus3 constellation with SA off, the availability of the RAIM detection function for a mask angle of 2.5 deg is compared with the case of a mask angle of 7.5 deg. In general, lowering the mask angle significantly improves availability for the RAIM detection and identification functions.

V. Summary and Conclusions

On the basis of the results, the following summary/conclusions are drawn.

1) Even if GPS is used as a supplemental system, high availability of the

Table 5 Availabilities for terminal and nonprecision approach phases of flight, 24Minus3 constellation, SA on, mask angle = 7.5 deg

	At 10,000 ft with protection limit of 1 n.mi.									At 1,500 ft with protection limit of 0.3 n.mi.								
	Navigation			Detection			Identification			Navigation			Detection			Identification		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
GPS alone																		
SFO	100	0	0	89.9	18	75	54.5	30	175	100	0	0	72.2	27	115	28.8	64	480
DFW	100	0	0	85.8	17	55	50.7	32	150	100	0	0	68.1	35	125	25.7	82	460
ORD	100	0	0	86.5	20	95	54.9	31	265	100	0	0	70.1	43	135	31.3	58	515
JFK	97.9	15	25	82.3	32	175	58.3	26	285	97	15	25	68.8	50	230	30.2	53	310
ATL	96.9	45	45	86.1	18	100	56.9	34	285	96.5	25	45	73.3	55	235	32.3	54	460
Baro aiding with GPS calibration in descending flight																		
SFO	100	0	0	92.7	26	75	80.0	32	145	100	0	0	72.6	30	115	42	70	445
DFW	100	0	0	90.3	23	55	77.4	22	135	100	0	0	69.1	34	125	36.8	70	460
ORD	100	0	0	92	29	95	73.3	35	200	100	0	0	71.5	46	135	43.8	68	355
JFK	100	0	0	87.2	46	140	76.7	34	235	99	8	10	69.1	40	225	46.5	39	285
ATL	100	0	0	90.3	35	100	77.1	33	255	99.7	5	5	74	63	235	46.5	51	335
Baro aiding with local pressure input																		
SFO	100	0	0	93.1	33	75	80.0	32	145	100	0	0	86.5	28	95	68.8	35	155
DFW	100	0	0	91.0	26	55	78.5	21	135	100	0	0	83.7	34	55	66	35	135
ORD	100	0	0	93.8	30	75	75.3	24	95	100	0	0	89.2	17	90	67	37	185
JFK	100	0	0	92.0	19	55	78.1	35	225	100	0	0	85.1	36	80	68.4	41	230
ATL	100	0	0	92.7	21	65	78.8	28	130	100	0	0	88.2	24	70	68.4	50	285
Clock coasting (1E-9 drift)																		
SFO	100	0	0	92.7	26	75	78.1	24	110	100	0	0	73.6	27	115	53.1	45	315
DFW	100	0	0	86.8	15	55	71.2	32	135	100	0	0	70.8	38	125	46.5	70	295
ORD	100	0	0	87.9	25	95	70.5	39	135	100	0	0	71.5	51	135	53.1	56	325
JFK	100	0	0	84.7	44	175	72.9	39	230	97.2	13	25	69.8	54	230	57.3	38	285
ATL	100	0	0	86.5	18	100	74	54	200	99.7	5	5	74.7	73	230	57.6	51	295

A = Availability, %.
B = Average outage duration, min.
C = Maximum outage duration, min.

Table 6 Availabilities of receiver autonomous integrity monitoring functions with and without SA, 24Minus3 constellation, mask angle = 7.5 deg

	Protection limit of 1 n.mi.						Protection limit of 2 n.mi.					
	SA on			SA off			SA on			SA off		
	A	B	C	A	B	C	A	B	C	A	B	C
	Detection function											
Baro aiding with GPS calibration in level flight												
JFK–LAX	93.9	48	80	94.6	43	75	94.9	40	75	95.2	38	70
SFO–NRT	94.3	31	85	95.4	30	80	95.7	20	50	96.6	22	50
DFW–CDG	96.3	16	50	97.6	19	50	98.1	15	35	98.6	15	35
Clock coasting (1E-9 drift)												
JFK–LAX	92.6	23	80	93.3	26	80	93.6	33	80	94.2	45	80
SFO–NRT	92.7	21	85	94.8	21	80	94.6	22	80	95.4	30	80
DFW–CDG	94.2	12	55	96.2	12	55	96.5	16	55	97.4	27	50
Combined use of baro and clock coasting												
JFK–LAX	93.9	48	80	94.6	43	75	94.9	40	75	95.2	38	70
SFO–NRT	94.1	32	85	95.4	30	80	96	22	50	96.6	22	50
DFW–CDG	96.3	16	50	97.6	15	50	98.2	14	35	98.7	20	35
	Identification function											
Baro aiding with GPS calibration in level flight												
JFK–LAX	80.4	24	100	86.9	21	95	86.5	19	95	89.1	17	95
SFO–NRT	83.3	20	145	87.7	19	125	87.2	18	125	90.1	18	125
DFW–CDG	88.1	17	125	91.2	15	95	91	14	75	94.2	13	70
Clock coasting (1E-9 drift)												
JFK–LAX	76.9	21	100	81.1	25	100	84	21	95	85.9	28	95
SFO–NRT	76.5	25	185	81.5	29	175	82.7	26	175	84.6	29	150
DFW–CDG	82.9	21	180	86.4	17	140	87.7	16	125	89.7	15	125
Combined use of baro and clock coasting												
JFK–LAX	85.3	21	95	89.1	24	95	91	28	95	91.3	27	95
SFO–NRT	86.4	24	145	89	22	125	91.4	19	125	92.7	20	120
DFW–CDG	91.5	16	95	93.6	14	85	94.7	12	70	95.7	15	70

A = Availability, %.
B = Average outage duration, min.
C = Maximum outage duration, min.

RAIM detection function during normal operation (e.g., with a typical 21-satellite constellation) is important to avoid arriving at the final approach fix and finding that a GPS approach cannot be conducted because the satellite geometry does not provide the RAIM detection function. This is undesirable especially if no other approach is available at the airport. Because barometric altimeter aiding increases the availability significantly, the FAA decided to require this augmentation in TSO-C129.

2) In the approach mode, barometric aiding with local pressure input improves availability significantly more than barometric altimeter aiding with GPS calibration. This led the SOIT to require use of local pressure information in the approach mode in TSO-C129.

Table 7 **Receiver autonomous integrity monitoring availabilities with two different mask angles, SA off, 24Minus3 constellation**

	Protection limit of 1 n.mi.						Protection limit of 2 n.mi.					
	7.5 deg			2.5 deg			7.5 deg			2.5 deg		
	A	B	C	A	B	C	A	B	C	A	B	C
	Detection function											
Baro aiding with GPS calibration in level flight												
JFK–LAX	94.6	43	75	96.5	55	55	95.2	38	70	96.8	25	45
SFO–NRT	95.4	30	80	99.5	15	15	96.6	22	50	99.7	10	10
DFW–CDG	97.6	19	50	99.8	5	5	98.6	15	35	99.8	5	5
Clock coasting (1E-9 drift)												
JFK–LAX	93.3	26	80	96.5	55	55	94.2	45	80	96.5	55	55
SFO–NRT	94.8	21	80	99.4	10	15	95.4	30	80	99.4	10	15
DFW–CDG	96.2	12	55	99.8	5	5	97.4	27	50	99.8	5	5
Combined use of baro and clock coasting												
JFK–LAX	94.6	43	75	96.8	25	45	95.2	38	70	97.4	40	40
SFO–NRT	95.4	30	80	99.5	15	15	96.6	22	50	99.7	10	10
DFW–CDG	97.6	15	50	99.8	5	5	98.7	20	35	99.8	5	5
	Identification function											
Baro aiding with GPS calibration in level flight												
JFK–LAX	86.9	21	95	92	21	70	89.1	17	95	92.9	16	70
SFO–NRT	87.7	19	125	97.5	10	20	90.1	18	125	98	9	20
DFW–CDG	91.2	15	95	98.2	14	30	94.2	13	70	98.9	9	20
Clock coasting (1E-9 drift)												
JFK–LAX	81.1	25	100	90.7	24	85	85.9	28	95	92.6	58	85
SFO–NRT	81.5	29	175	96.6	16	50	84.6	29	150	97.4	14	50
DFW–CDG	86.4	17	140	96	14	45	89.7	15	125	97.4	20	40
Combined use of baro and clock coasting												
JFK–LAX	89.1	24	95	93.9	24	70	91.3	27	95	95.2	38	70
SFO–NRT	89	22	125	98.6	9	20	92.7	20	120	98.9	9	20
DFW–CDG	93.6	14	85	98.9	12	20	95.7	15	70	99.8	5	5

A = Availability, %.
B = Average outage duration, min.
C = Maximum outage duration, min.

3) However, even with the use of local pressure input, the availability of the detection function at the final approach fix is only 84–89% with the 24Minus3 constellation. According to FAA certification specialists, in order for a navigation system to be useable for nonprecision approach, the minimum availability of approach capability (which requires the RAIM detection function to exist) upon arrival should be at least 95%. Because this goal cannot be achieved, even with use of local pressure input for the 24Minus3 constellation, and because the majority of outage periods are predictable once the constellation is known, TSO-C129 requires GPS receivers to be able to predict RAIM availability at the destination airport at the estimated time of arrival.

4) For the other phases of flight (i.e., oceanic, en route, and terminal), the availability of the RAIM detection function is high enough (typically 95% or higher with the 24Minus3 constellation) so that GPS may successfully be used as a supplemental system. In order for GPS to be used as a sole-means (standalone) system, however, a very high availability (close to 100%) is required both for detection and identification functions. However, the analysis indicated that availability of the identification function is not high enough even for the oceanic phase of flight (i.e., about 95% with the 24Minus3 constellation). Therefore, for GPS is to be used as a sole-means system, a system such as an inertial system, or some other augmentation would be required.

5) Clock coasting with a user receiver clock with a drift of 10^{-9} is not as effective as barometric altimeter aiding, especially in level flight. As stated earlier, however, if a microminiature atomic clock with a drift of 10^{-11} or smaller becomes available in the future, clock coasting would significantly improve RAIM availability. Also, estimation of the clock rate may be promising, although the presence of SA would limit the accuracy with which the clock rate can be estimated.

Appendix: Statistical Distribution of the Height Gradients[†]

A. Background

An aircraft is assigned a flight level, which means it follows whatever height gives a constant reading on its barometric altimeter. A particular law of height vs pressure is then assumed in order to derive height. Even if the pilot maintains a constant height reading of his barometric altimeter, the aircraft is for most of the time actually altering its vertical distance from mean sea level because of deviations from the assumed pressure–height law and variations in true ground pressure.

This becomes important for the process of aiding the GPS receiver, because the height input the receiver can utilize is the height indicated by the barometric altimeter. Because that height is in error, it follows that there will be an error in the receiver's output. Three factors have an impact on the likely success in aiding the receiver from a barometric altimeter: 1) the rate at which this altitude error varies with distance flown; 2) the characteristics of the error arising when the GPS receiver with its erroneous height is used to calibrate the barometric altimeter; and 3) the geometric limitation inherent in treating the information from the barometric altimeter as a distance from the Earth's center (i.e., the dilution of precision when this sensor is used).

Feature 1) is specific to the atmosphere and has nothing to do with GPS. Feature 2) is specific to GPS and has nothing to do with the atmosphere, but is a result of range error distributions modified by GPS geometry. Feature 3) is a combination of GPS and baormetrically measured altitude; whereas the satellite moves about, altitude is always relative to the Earth's center. In each of these cases, we could work out a probability distribution for the consequences.

[†] R. Johannessen and C. Dixon, BNR Europe Limited, Harlow, England.

B. Technique Adopted

Meteorological measurements from aircraft and radiosonde measurements are fed into the model[16] developed by the UK Meteorological (MET) Office, which produces, among other parameters, values for height having a given barometric pressure, at fine intervals in latitude and longitude. This is done at regular intervals of time as part of the work of the Met Office. A selection of these values were stored on floppy disks and given to BNR Europe Limited, in Harlow, England for analysis, funded from the Chief Scientist's Division of the UK CAA in London. Geopotential height values were provided for both 250 mbar and for 850 mbar, corresponding approximately to heights of 33,000 ft and 4000 ft, respectively. These height values on a regular geographic interval, taking account of the horizontal separation between the sampling points, provided a large number of slopes expressed in meters height change per nautical miles horizontal distance flown when maintaining constant barometric pressure.

Samples were taken between 80.625W and 30E at intervals of 0.9375° long., and between 30N and 60N at intervals of 0.75° lat. 4879 points were available. At each of these points four gradients were computed so that some 19,000 gradients resulted for each day. Twelve different days (the first day of every month for one year) were sampled providing a data base of 225,000 gradients, all at 250 mbar. Additionally a subset of points were chosen at 850 mbar. These individually computed values for slope were then analyzed statistically and the probability distribution derived.

C. Results Obtained (Table A.1)

Taking a typical aircraft speed of 450 kt and assuming an integrity outage time of 10 min, the aircraft will have traveled 75 n.mi. leading to a height variation of less than $(0.92 \times 75) = 69$ m at 95% level of confidence (250 mbar). Likewise, if the integrity outage is 60 min, the corresponding height variation is 414 m. In this context, it should be noted that at least six satellites are needed to provide fault isolation, whereas at least five are needed for fault detection. The proportion of time that the requirements for fault isolation are met is, therefore, lower than the proportion of time that the requirements for fault detection are met. It follows that the duration of the periods when a receiver cannot isolate a faulty satellite will be longer than the duration of periods when a receiver cannot detect whether or not the integrity is good. Whereas the 69 m above may be appropriate for fault detection difficulties (as is of interest to supplemental means navigation), the longer 414 m is likely to be more relevant to the harder fault isolation cases (as is important to sole means navigation).

For comparison, it is noted that Ref. 17 has 0.7 m/n.mi. as "typical" for 40,000 ft altitude over the United States and that Ref. 10 has 1.0 m/n.mi. for 30,000 ft altitude, also over the United States. It seems from the latter paper that the 1.0 m/n.mi. is derived by looking at the most closely spaced pressure contours, thus providing a worst case value. Both these reference therefore, are, in broad agreement with Table A.1.

The probability distribution in Table A.1 can be used in two different ways. First, we can take a simulated flight proceeding along the track until there is an "outage"; i.e., when the RAIM or navigation functions cannot be performed

Table A.1 Height change/distance flown for different probabilities at 250 and 850 mbar[a]

Confidence level, %	Slope in m/n.mi.	
	250 mbar	850 mbar
5	0.04	0.04
15	0.07	0.04
25	0.13	0.06
35	0.18	0.07
45	0.23	0.09
55	0.31	0.13
65	0.38	0.16
75	0.49	0.20
85	0.65	0.25
95	0.92	0.36

[a] 250 mbar and 850 mbar correspond approximately to heights of 33,000 ft and 4000 ft, respectively.

without aiding. At that stage, the barometric height is "calibrated," and the navigation or RAIM functions continue in height aiding mode with the height error changing because of the slope. Both the calibration error and the slope error are selected at random; the former from a distribution curve for height error in GPS and the latter from the distribution of Table A.1. This will most closely represent the procedure followed in the navigating aircraft.

Second, we can approximate the slope distribution to a standard deviation, and combine that error along with the pseudorange errors through normal matrix principles. The distribution was derived in order to allow the first option. On the other hand, the analysis contained earlier in the chapter used the second option.

References

[1] Lee, Y., "RAIM Availability for GPS Augmented with Barometric Altimeter Aiding and Clock Coasting," *Navigation,* Vol. 40, No. 2, 1993.

[2] "Airborne Supplemental Navigation Equipment Using the Global Positioning System (GPS)," Technical Standards Order (TSO)-C129, FAA Aircraft Certification Service, Washington, DC, Dec. 10, 1992.

[3] Brown, A., and Schmid, T., "Integrity Monitoring of the Global Positioning System Using a Barometric Altimeter," *Proceedings of the Institute of Navigation National Technical Meeting,* (Santa Barbara, CA), Institute of Navigation, Washington, DC, Jan. 26–29, 1988.

[4] Brown, R., Grover, G., Chin, Y., and Kraemer, J. H., "Update on GPS Integrity Requirements of the RTCA MOPS," *Proceedings of the Institute of Navigation Satellite Division's International Technical Meeting,* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 11–13, 1991 (ION GPS-91).

[5] Van Dyke, K., "RAIM Availability for Supplemental GPS Navigation," *Navigation,* Vol. 39, No. 4, 1992–1993, pp. 429–43.

[6] Brown, R. G., and McBurney, P., "Self-Contained Integrity Check Using Maximum Solution Separation as the Test Statistic," *Navigation,* Vol. 35, No. 1, 1988, pp. 41–54.

[7]Lee, Y., "Performance Analysis of Self-Contained Methods for GPS Integrity Function," The MITRE Corporation, MTR-88W89, McLean, VA, Nov. 1988.

[8]"Minimum Operational Performance Standards for Airborne Supplemental Navigation Equipment Using Global Positioning System (GPS)," Radio Technical Commission for Aeronautics RTCA DO-208, Washington, DC, July 1991.

[9]Asbury, M. J. A., Forrester, D. A., Dixon, C. S., and Johannessen, R., "Probability Distributions That are Important When Assessing Barometric Aiding to GPS," *Proceedings of the Institute of Navigation Satellite Division's International Technical Meeting,* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 19–21, 1990 (ION GPS-90).

[10]Dobyne, J., "The Accuracy of Barometric Altimeters with Respect to Geometric Altitude," *Proceedings of the Institute of Navigation Satellite Division's International Technical Meeting,* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 19–23, 1988.

[11]Brown, G., "A Baseline RAIM Scheme and a Note on the Equivalence of Three RAIM Methods," *Navigation,* Vol. 39, No. 3, 1992, pp. 301–316.

[12]Lee, Y., "Receiver Autonomous Integrity Monitoring (RAIM) Capability for Sole-Means GPS Navigation in the Oceanic Phase of Flight," *IEEE Aerospace and Electronic Systems Magazine,* Vol. 7, No. 5, 1992, pp. 29–36.

[13]Winkler, G., "Briefing Presented to the DOD GPS Integrity Technical Tiger Team Meeting," Colorado Springs, CO, May 21, 1992.

[14]Green, G., et al., "The GPS 21 Primary Satellite Constellation," *Navigation,* Vol. 36, No. 1, 1989, pp. 9–24.

[15]Dobyne, J., "GPS Availability," Paper presented to the FAA Satellite Operational Implementation Team, Dec. 10–12, 1991.

[16]Bell, R. S., and Dickinson, A., "The Meteorological Office Operational Numerical Weather Prediction System," Met Office Scientific Paper 41, HMSO, London, 1987.

[17]Brown, A., "Integrity Monitoring of GPS using a Barometric Altimeter," (Preliminary report), RTCA Paper 405-87/SC159-117.

Chapter 9

GPS and Global Navigation Satellite System (GLONASS)

Peter Daly*
University of Leeds,
Leeds, LS2 9JT, England, United Kingdom
and
Pratap N. Misra†
Massachusetts Institute of Technology,
Lexington, Massachusetts 02173

I. Introduction to the Global Navigation Satellite System

A. History of Satellite Navigation Systems

BOTH the NAVSTAR Global Positioning System (GPS) and Global Navigation Satellite System (GLONASS) developed respectively by the United States and the (former) Union of Soviet Socialist Republics (USSR), now the Commonwealth of Independent States (CIS), are planned to become operational during the 1995–1996 time period. Known under the generic title of Global Navigation Satellite Systems (GNSS), they are intended to replace earlier satellite navigation systems (Transit and Cicada) also operated by the United States and the USSR. These two latter systems employ similar orbits with a small number of low-altitude (1100 km) polar-orbiting satellites transmitting information at dual frequencies around 150 and 400 MHz. The user waits for a single satellite (possibly as long as 2 h) and then makes a series of Doppler measurements during the short period (<16 min) when the satellite remains above the horizon. The satellites' position and velocity are transmitted in the navigation message and these, together with the Doppler measurements, are sufficient to allow the user to compute a position. Transmissions on two frequencies are used to allow an ionospheric group delay correction to be applied. The two systems have two major drawbacks; they are not available continuously and the user velocity must be known.

In an effort to overcome the difficulties associated with the earlier systems, both the United States and the CIS plan to introduce precise, global, continuous

*Director, CAA Institute of Satellite Navigation.

†Senior Staff Member, Lincoln Laboratory.

position-fixing capabilities by using navigation satellites transmitting dual-frequency spread-spectrum signals in L-band (1.2 and 1.6 GHz). In contrast to the earlier VHF systems, the primary navigation mode is based on range measurement rather than integrated Doppler. The two national systems, both of which possess a military and a civil role, are the CIS's GLONASS and the USA's NAVSTAR GPS,[1] designed to provide accurate position, velocity, and time information. At the end of 1993, the Initial Operating Capability (IOC) had been declared for GPS, while GLONASS was still in the preoperational stage, with a number of satellites already performing a navigation role.

The first release from the Soviet Union of detailed GLONASS information occurred at the International Civil Aviation Organization (ICAO) special committee meeting on Future Air Navigation Systems (FANS) in Montreal in May 1988.[2] At a later meeting of ICAO in September 1991, the concept of GNSS was adopted as providing for future air navigation requirements. The notion of GNSS encompasses GPS, GLONASS, and such alternative systems as the geostationary overlay to be provided by Inmarsat-3 satellites. GLONASS is intended to provide a navigation role for maritime and aviation interests; it offers many features in common with NAVSTAR GPS. In particular, the orbital plan foresees 24 satellites with 8 in each of three orbital planes separated by 120 deg with spacing of 45 deg within the plane. Clearly the orbital planning is such as to allow users anywhere access to at least the minimum number of satellites (four) required for navigation purposes. In practice, simple geometrical considerations tell us that, when fully operational, both GPS and GLONASS individually will allow access to eight or nine satellites for the greater part of each day. The combined resources of GPS and GLONASS together offer twice as many satellites as either system taken on its own. This doubling of available satellite numbers offers a level of independence, redundancy and cross-checking that enhances certain global applications of GNSS such as civil aviation, as discussed in detail in Sec. II.C.

GLONASS transmits two spread-spectrum signals in L-band at around the same power levels as NAVSTAR GPS. Satellites are distinguished by radio frequency channel rather than spread-spectrum code. A single narrow-band code is used of length 511 bits repeating every 1 ms. Information is differentially encoded in a return-to-zero (RZ) format with a final datarate of 50 baud. A separate broad-band code repeating every second is used to transmit differentially encoded data at 50 baud. The narrow-band code is to be found only at the higher of the two L-band transmit frequencies; the broad-band code is to be found on both the upper and lower L-band frequencies, and hence, offers the prospect of correction for the ionospheric delay effect. In this regard, the situation is entirely analogous to that of GPS. A plot of a typical GLONASS signal spectrum is shown in Fig. 1. The GLONASS C/A-code (coarse/acquisition) spectrum covering 1 MHz bandwidth is superimposed on the P-code (precise) signal transmitted in phase quadrature and covering 10 times the C/A-code bandwidth. On the ground, the spectrum is only reproducible in the first place by using a high-gain antenna (3-m dish or larger) to extract the spread-spectrum signal from the noisy background.

B. Orbits

For a given number of satellites in the final operational system, the choice of orbital planes and phases within the plane is constrained to ensure visibility of

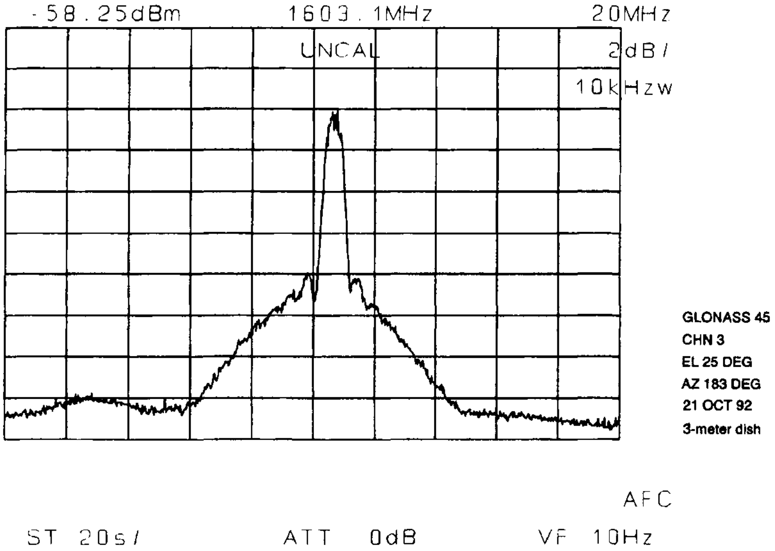


Fig. 1 Plot of typical GLONASS signal spectrum.

four well-located satellites on a continuous global basis and graceful degradation of the system during spacecraft failures. A common approach is to adopt a small number (three or six) equally separated inclined orbital planes with a number of satellites distributed equally in phase around each plane and with an offset in phase between planes. It is also possible to augment this approach with a number of satellites in the geostationary arc. The GLONASS satellite navigation system foresees an operational configuration of 24 satellites with 8 satellites in each of three orbital planes separated by 120 deg in Right Ascension of the Ascending Node (RAAN). RAAN may be thought of simply as equator-crossing longitude but with reference to an inertial (star-fixed) frame rather than a rotating (Earth-fixed) frame. The situation as of September 1994 is shown in Fig. 2 showing a separation in argument of latitude or orbital phase in the plane of 45 deg. There is also a displacement of +30 deg and −30 deg for satellites in planes two and three, respectively, with reference to plane one. This nomenclature follows that assumed by the GLONASS almanac format (described later in Sec. I.G). Relative positions of satellites remain very stable over long periods because they have very much the same, small rates of change of RAAN with time amounting to about −0.03 deg/day for near-circular GLONASS orbits.

All satellites have the same nominal orbital period of 675.73 min with longitude change of 169.41° W each orbit. This orbit produces a ground-track repeat every 17 orbits lasting 8 whole days less 32.56 min. The diurnal offset of $\Delta T = 4.07$ min from a full 24-h day coincides with that of NAVSTAR GPS and is very nearly the difference between a solar and sidereal day (3.93 min). This implies that each complete day less ΔT minutes a satellite performs 17/8 orbits or two whole revolutions plus an additional 1/8 revolution, equivalent to 45 deg. It follows that two satellites in the same plane but separated by 45 deg in orbital

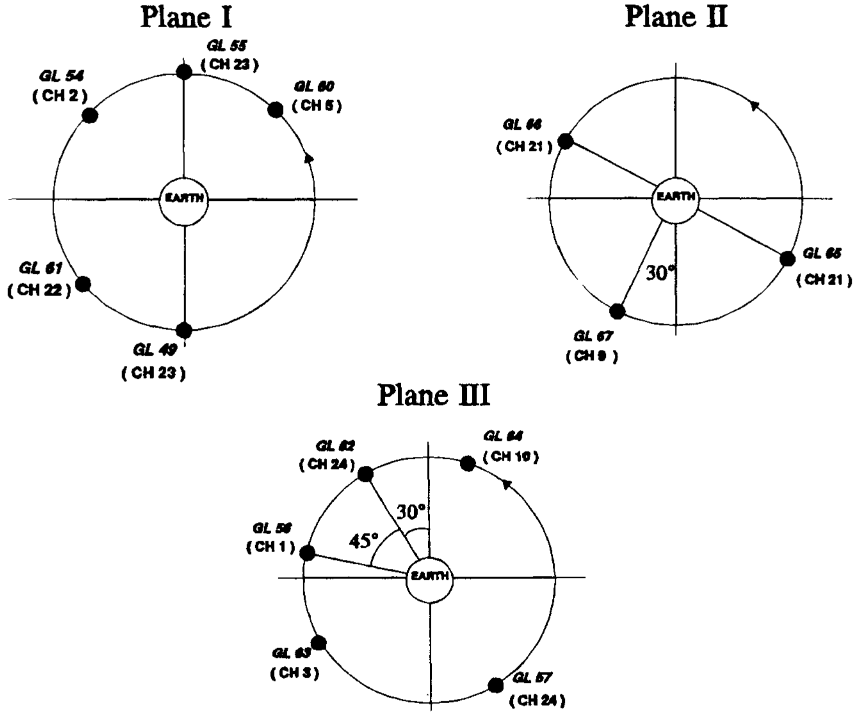


Fig. 2 GLONASS operational configuration as of September 1994.

phase, appear at precisely the same position on successive days less ΔT minutes. During that interval, the Earth has rotated very nearly 360 deg with the result that the ground-based observer sees both satellites at the same pointing azimuth and elevation. Over a ground-track repeat interval of eight days, then, all satellites in the same plane with separation of 45 deg appear in turn at the same position at intervals of 1 day less ΔT minutes. After 8 days, the whole cycle naturally repeats.

By examining the phases of satellites in the planes two and three, it becomes apparent that these satellites will also appear at the same position as the reference satellite in plane one within the same 8-day period. This arises because the time taken by the Earth to rotate through the angle 120 deg separating planes one and two is the same time taken by a satellite in that plane with phase +255 deg to travel around to the same position as the reference satellite. The Earth rotates through 120 deg in 478.69 min, very nearly 8 h, which corresponds almost exactly to 17/24 of a GLONASS orbit or +255 deg. The same argument holds for plane three at 240 deg separation for a satellite at phase +150 deg (or twice +255 less 360 deg). The angular separation of 45 deg within the plane together with the angular phase differences of ± 30 deg between planes assures that in an 8-day period, all 24 satellites will pass through the position with the reference subsatellite location.

C. History of Launches

Global satellite navigation systems have been under development by the United States and the former Soviet Union since the 1970s, although the nations already operated and, in fact, still operate a dual-frequency VHF satellite navigation system in near-polar, low Earth-orbit (U.S. Transit, USSR Cicada) providing limited time access but on a worldwide basis. The USA NAVSTAR GPS system saw its first launch in 1978; the USSR's GLONASS system was inaugurated 4 years later. GLONASS satellites are launched three at a time from the Tyuratam space center into near-circular orbits with period around 11.25 h at a height of 19,100 km and inclination of 64.8 deg. Before the demise of the Soviet Union, successful launches were followed by an announcement within a day or two in *Pravda,* which printed basic details of the mission. As of September 1994, there have been 23 launches since the first one in late 1982, all of them successful except those of April 1987 and February 1988, wherein the satellites failed to reach final orbit because of a malfunction of the fourth stage of the Proton launch vehicle. All launches but the most recent (launches 19–23) have taken place under the auspices of the USSR. Since the demise of the Union in 1991, the launches that took place were conducted by the CIS. Table 1 presents the international identifiers, Cosmos and GLONASS numbers of all known launches since

Table 1 Triple GLONASS launches history, September 1994

	International designator	Cosmos	GLONASS	Plane
1	1982—100A,D,E	1413 1414 1415	1 2 3	1
2	1983—84A,B,C	1490 1491 1492	4 5 6	1
3	1983—127A,B,C	1519 1520 1521	7 8 9	3
4	1984—47A,B,C	1554 1555 1556	10 11 12	3
5	1984—95A,B,C	1593 1594 1595	13 14 15	1
6	1985—37A,B,C	1650 1651 1652	16 17 18	1
7	1985—118A,B,C,	1710 1711 1712	19 20 21	3
8	1986—71A,B,C	1778 1779 1780	22 23 24	1
9	1987—36A,B,C	1838 1839 1840	25 26 27	3[a]
10	1987—79A,B,C	1883 1884 1885	28 29 30	3
11	1988—9A	1917 1918 1919	31 32 33	1[a]
12	1988—43A,B,C	1946 1947 1948	34 35 36	1
13	1988—85A,B,C	1970 1971 1972	37 38 39	3
14	1989—1A,B,C	1987 1988 1989	40 41[b]	1
15	1989—39A,B,C	2022 2023 2024	42 43[b]	3
16	1990—45A,B,C	2079 2080 2081	44 45 46	3
17	1990—110A,B,C	2109 2110 2111	47 48 49	1
18	1991—25A,B,C	2139 2140 2141	50 51 52	3
19	1992—5A,B,C	2177 2178 2179	53 54 55	1
20	1992—47A,B,C	2204 2205 2206	56 57 58	3
21	1993—10A,B,C	2234 2235 2236	59 60 61	1
22	1994—21A,B,C	2275 2276 2277	62 63 64	3
23	1994—50A,B,C	2287 2288 2289	65 66 67	2

[a] Failed to achieve final orbit.

[b] Etalon passive laser ranging satellite.

the first one of 1982; all triple launches, except the most recent, have taken place into one of two orbital planes (referred to in the table as planes 1 and 3) separated by 120 deg. The most significant novel feature of the most recent launch is that for the first time, plane 2 was chosen. The decision to begin to occupy the third plane can be interpreted as indicating serious intent to complete the entire 24-satellite GLONASS constellation.

Each launch aims to produce a final stable near-circular inclined orbit at a distance from the Earth's center of about four Earth radii. Of the first seven launches only two of the three launched satellites achieved the said stable orbit; the third satellite remained in an orbit without ground-track repeat and was not observed to transmit. Since then, launches 8, 10, 12–13, and 16–23 have all resulted in a stable orbit for all three satellites. They have also transmitted full navigation messages and can, therefore, be regarded as full-fledged members of the preoperational system. Launches 9 and 11 were failures caused by a malfunction of the fourth stage of the Proton SL-2 launcher. Launches 14 and 15 placed two (rather than the normal three) GLONASS satellites into stable orbit. In both cases, a third satellite was launched with the two GLONASS spacecraft. This third spacecraft (known as ETALON) was a passive spheroidal satellite provided with reflectors to act as a target for laser-ranging signals.

It was also a normal feature of the system in its early days for a launch to occur only when the number of functioning satellites had fallen or was about to fall below the number required for adequate testing of the system. This number cannot be stated with absolute precision because it depended on the orbital planes of the remaining satellites. However, a reduction of available units to any number less than four was likely to act as a precursor to a new launch. Given the number of satellites launched since the first one of 1982, we must presume that these have taken place to make up for the substantial number of in-orbit failures in the interim period. The attrition rate for GLONASS satellites is, indeed, very high (as compared with GPS), although it is not as easy as we might imagine for Western sources to determine whether or not a particular satellite has failed. On the face of it, the failure of a satellite is apparent when the satellite no longer transmits a navigation signal. However, it has been observed in the past that a satellite can remain in a passive, stand-by mode for 3 months before being returned to service. It is normally (but not always) quite correct to assume a satellite has failed when signals are no longer received over a period of several weeks.

D. Signal Design

A major aim of GLONASS is to provide high-precision position-fixing and time-reference capability on a worldwide and continuous basis to users on the Earth's surface, on land and at sea, in the air, and in space itself. Under the control of highly stable, onboard frequency references (clocks), timing signals (epochs), and data are transmitted simultaneously from a number of satellites. The data received from a particular satellite includes a precise ephemeris for that satellite allowing its position and velocity at a given time to be computed. In addition, each satellite provides information on the behavior of its own onboard clock. The observer makes time-of-arrival measurements to three satellites at the same time using the received data to compute the position of the satellites. Given a synchronized ground time reference, the observer can then, in principle, solve three range equations for three unknown position coordinates. In practice, the observer will not normally have a synchronized time reference and will, therefore, choose to "pseudorange" to four satellites instead of three, using the fourth measurement to compute the instantaneous time error.

Users of satellite navigation have the option to make measurements at the level of code phase or to enhance this with integration of Doppler frequency for averaging purposes or when in motion. In addition, the user may have access to carrier phase, which allows extremely accurate relative position fixing when used[3] in the differential mode. This particular method is of great use to the surveying community. In addition, a technique known as kinematic positioning has been derived[4] based on the notion of relative positioning between two observers, one of which is in motion, continuous, or in steps.

For the purposes of allowing computation of the user's own position, navigation satellites transmit details of their own positions and a time reference. In systems such as GLONASS and NAVSTAR GPS, whose purpose is primarily military, the user is expected to play a passive role, because any transmissions to a satellite might identify the user's position to an adversary. Similarly, the navigation message is protected against deliberate jamming by the use of spread-spectrum codes, which increases the bandwidth occupied by the signal and, hence, that of the intending jammer. It should be clear, however, that even were the system design to be exclusively for civil purposes, it is likely that spread spectrum would still be used for two reasons: 1) to minimize interference to others; and 2) to provide sufficient bandwidth and, hence, definition of the epoch timing edge. In the following discussion, the structure of the navigation signal from global navigation satellites to the user is treated.

The transmission carrier frequencies chosen for the new satellite navigation systems lie in L-band, around 1250 MHz (L_2) and 1600 MHz (L_1). Dual-frequency navigation transmissions at L_1 and L_2 allow the user to correct for ionospheric propagation effects and are incorporated into both NAVSTAR and GLONASS. A high-precision spread-spectrum code is modulated onto both carriers; whereas, the lower-precision civil code only appears at L_1. Russian publications on this subject often refer to the low-precision codes as providing the GLONASS SPS (standard positioning service) and to the higher-precision codes as providing the GLONASS "M" PPS (precise positioning service). Spread-spectrum techniques are primarily involved to reduce the effects of deliberate jamming of signals.

In stark contrast to the radio frequency transmission system chosen for GLONASS, which is FDM (frequency division multiplex), NAVSTAR GPS uses the same carrier frequency for all satellites that are distinguished one from another by the use of different spread-spectrum Gold codes, a form of CDM (code division multiplex). This difference between the two systems is of major significance in designing receivers capable of joint operation. It is perhaps worth noting in passing that a third global satellite navigation program proposed by the European Space Agency (ESA Navsat) is characterized by the use of TDM (time division multiplex) to distinguish individual satellites.

The GPS satellites use transmit frequencies at L_1 (1575.42 MHz) and L_2 (1227.60 MHz). Radio frequency carriers used by GLONASS are channelized within the bands 1240–1260 MHz and 1597–1617 MHz, the channel spacing being 7/16 or 0.4375 MHz at the lower frequencies and 9/16 or 0.5625 MHz at the higher frequencies. The carrier frequencies themselves are also multiples of channel spacing and the number of planned channels is 24. The relationship between channel number and the L_1 transmit frequency is simply $f(L_1) = f_0 + k \times$ channel spacing, where $f_0 = 1602$ MHz, and k is the channel number (1–24). Transmit frequencies at the lower (L_2) frequencies are in the ratio 7/9. GLONASS satellites and channel numbers (designated GL and CHN) are presented in Table 2. All frequencies were measured at the University of Leeds, with the exception of that of GLONASS 5, which was obtained at the Haystack Laboratory in the United States. At the time of writing (Sept. 1994), there are 13 active GLONASS satellites. Note that toward the end of September 1993 following the regulations introduced at WARC 92, the technique of using the same channel for certain antipodal satellites was introduced.

As explained in a previous section, each satellite sends data at low speed from which its own position at any reference time may be calculated. This data, commonly sent at a 50-baud rate, is superimposed on a pseudorandom noise (PRN) code that is, in fact, periodic and very much longer than a single databit. The GLONASS low-precision code has length 511 bits as compared to NAVSTAR's 1023 bits for its equivalent code. A code sequence lasts only 1 ms, so each databit occupies 20 entire code sequences, the code itself or its inverse being sent depending on whether the databit is a "0" or a "1." In this manner, the information spectrum is spread across a wide bandwidth. On the assumption that, in transmission, the signal will be corrupted by Gaussian noise whose power level is proportional to bandwidth, then the signal will become immersed in the noise at the receiver's terminal and recoverable only be reversing the coding operation applied at the transmitter. This implies a knowledge of the PRN codes on the part of the receiver.

The code rate can be seen from the numbers already given to be 511 kbits/s and 1023 kbits/s for the civil GLONASS and NAVSTAR codes, respectively. Higher-rate codes are at 10 times these rates, and, of course, the sequence lengths are very much longer. To transmit the encoded data, a binary phase-shift keyed (BPSK) modulation technique is employed, the first nulls in the transmitted spectrum at ± the bit rate. Hence, bandwidths for the GLONASS transmission can be taken at 1 MHz and 10 MHz for the low-rate (C/A) and high-rate (P) codes, respectively. These figures compare with 2 MHz and 20 MHz for NAVSTAR's equivalent bandwidths. Both GLONASS codes are generated as maximum-length sequences. The corresponding generator polynomials are

C/A-code generator

$$g(X) = 1 + X^5 + X^9$$

and

P-code generator

$$g(X) = 1 + X^3 + X^{25}$$

the latter code[5] being truncated after every second. There is no need in GLONASS for a hand-over word (HOW) as with GPS to allow acquisition of the P-code. At the L_2 frequency, only the high-rate code is carried, but at L_1 both codes are transmitted on the same carrier, one in-phase and the other in-quadrature. This

Table 2 GLONASS channels

Sat ID	Cosmos	GL	Active	CHN
1982—100A	1413	1	——[a]	——[a]
1982—100D	1414	2	——[a]	——[a]
1982—100E	1415	3	——[a]	——[a]
1983—84A	1490	4	N	3
1983—84B	1491	5	N	1
1983—84C	1492	6	——[a]	——[a]
1983—127A	1519	7	N	2
1983—127B	1520	8	N	24
1983—127C	1521	9	——[a]	——[a]
1984—47A	1554	10	N	9
1984—47B	1555	11	N	18
1984—47C	1556	12	——[a]	——[a]
1984—95A	1593	13	——[a]	——[a]
1984—95B	1594	14	——[a]	——[a]
1984—95C	1595	15	N	17
1985—37A	1650	16	N	7
1985—37B	1651	17	N	10
1985—37C	1652	18	——[a]	——[a]
1985—118A	1710	19	N	4
1985—118B	1711	20	N	19
1985—118C	1712	21	——[a]	——[a]
1986—71A	1778	22	N	11
1986—71B	1779	23	N	20
1986—71C	1780	24	N	22
1987—79A	1883	28	N	14
1987—79B	1884	29	N	21
1987—79C	1885	30	N	5
1988—43A	1946	34	N	12
1988—43B	1947	35	N	23
1988—43C	1948	36	N	24
1988—85A	1970	37	N	18
1988—85B	1971	38	N	7
1988—85C	1972	39	N	10
1989—1A	1987	40	N	9
1989—1B	1988	41	N	6
1989—39A	2022	42	N	16
1989—39B	2023	43	N	17
1990—45A	2079	44	N	21
1990—45B	2080	45	N	3
1990—45C	2081	46	N	15
1990—110A	2109	47	N	4
1990—110B	2110	48	N	13
1990—110C	2111	49	Y	19
1991—25A	2139	50	N	20
1991—25B	2140	51	N	11
1991—25C	2141	52	N	14
1992—5A	2177	53	N	22
1992—5B	2178	54	Y	2

[a] Not yet observed.

(Continued on next page)

Table 2 GLONASS channels (continued)

Sat ID	Cosmos	GL	Active	CHN
1992—5C	2179	55	Y	17
1992—47A	2204	56	Y	1
1992—47B	2205	57	Y	24
1992—47C	2206	58	N	8
1993—10A	2234	59	N	12
1993—10B	2235	60	Y	5
1993—10C	2236	61	Y	23
1994—21A	2275	62	Y	24
1994—21B	2276	63	Y	3
1994—21C	2277	64	Y	10
1994—50A	2287	65	Y	21
1994—50A	2288	66	Y	21
1994—50C	2289	67	Y	9

results in a signal spectrum that superimposes the two individual spectra whose bandwidths differ by a factor of 10. Because both transmissions carry roughly the same power, a spectrum analyzer will display the narrower-band code at 10 times the strength of the wide-band code.

E. Message Content and Format

The data carried on transmissions from satellites are low bit rate at 50 baud essentially providing accurate positions for the transmitting satellite as well as information on its onboard frequency standard. In addition, data are given in the form of low-precision almanacs of all the other satellites currently available so as to allow the user to plan usage and to assist with signal acquisition. Data are sent in lines, subframes, and frames with preambles and parity checks at the end of each line. Essentially, each subframe of data lasts for 30 s and consists of 15 lines of 2-s duration. The 15-line subframe divides into a 5-line set of ephemeris data (including clock corrections) and a 10-line set of almanac data. The almanac data are subcommutated, each satellite being allocated 2 lines of almanac data. In this way, the subframe accommodates the almanacs of five satellites; the full constellation of 24 satellites occupies five complete subframes, the last of the 25 almanacs always being set to zero (spare). The subframe format is shown in Fig. 3 (the number of bits allocated to each parameter is found under the title of the parameter).

The start of each line marks the beginning of a 2-s subframe synchronized to GLONASS system time. The leading bit is always zero and followed by a line number, various databits, parity bits, and preamble. Some of the information bits are flags (for example, words *P*1-*P*5) and are not used by the navigator. The flag *H* on line 2 is most important, referring to the "health" of the space vehicle ("1" means unhealthy). The symbol *TA* gives the time of transmission in hours, minutes, and half-minutes Moscow Standard Time [MST (GMT + 3 h)]. The symbol *TE* gives the time of validity of the ephemeris in hours and quarter-hours

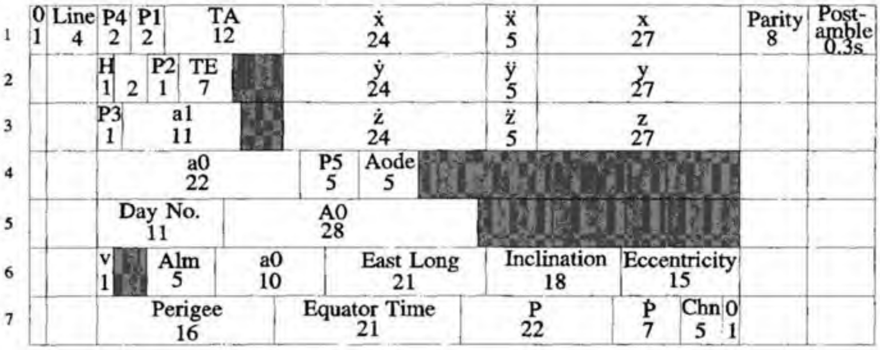


Fig. 3 GLONASS C/A-code data message subframe format.

(MST). Usually the ephemeris is valid at odd quarter-hours, but occasionally (when a satellite is temporarily unhealthy) the ephemeris will be valid only on the half-hour. The remaining databits are described in the following sections. The reader is referred elsewhere for further details of the preamble and parity corrections.[6]

F. Satellite Ephemerides

The NAVSTAR GPS ephemeris data describe a Kepler ellipse with additional corrections that then allow the satellite's position to be calculated in an Earth-centered, Earth-fixed (ECEF) set of rectangular coordinates at any time during the period of validity of the data. In the GLONASS system, the satellite's instantaneous position and velocity are encoded at fixed time intervals (usually one half-hour) in an ECEF rectangular coordinate system. Positions and velocities at intermediate times are intended to be calculated using interpolation procedures and acceleration terms provided. The resolution in satellite position is 0.5 m and in velocity, 1 mm/s.

As far as timing is concerned, the global navigation satellites transmit satellite clock corrections to a GPS or GLONASS system time and, in addition, corrections from system time to a national time reference. NAVSTAR GPS represents satellite clock behavior in terms of a clock offset ($a0$), frequency offset ($a1$), and rate of change of frequency ($a2$). NAVSTAR GPS system time from Universal Coordinated Time at the U.S. Naval Observatory, UTC(USNO), is given as $A0$, and its first derivative as $A1$. In contrast, GLONASS only transmits in the ephemeris two parameters relating to the onboard clocks, $a0$ and $a1$. The first time offset (with resolution 1 ns) refers to the instantaneous time difference between space vehicle time and GLONASS system time. The second parameter (with resolution 1 in 10^{-12}) gives the rate of change of space vehicle time offset. Use of both parameters allows the user to establish individual space vehicle time offsets at any required instant of observation. Reference of GLONASS system time to Universal Coordinate Time/Soviet Union, UTC(SU) is by a single time offset parameter, $A0$ (with resolution 7 ns). An additional parameter called age-of-data of ephemeris (AODE) gives the integer number of days starting at the previous local midnight since the ephemeris data was updated. On a particular day, the parameter will normally be "0," increasing by "1" each successive midnight unless, in the interim, the ephemeris data are based on a current (same day) set of measurements. When the GLONASS system is operating normally, this parameter is always either "0" or "1". On occasion it has been observed to grow as large as 25.

G. Satellite Almanacs

There is greater similarity between NAVSTAR and GLONASS in the transmission of almanacs than in the transmission of ephemerides. Both systems transmit the basic elements of an osculating Kepler ellipse as Table 3 attempts to show.

In terms of using almanacs to predict satellite position from the reduced Kepler orbit, the two sets of data are quite similar, as we would expect.

Each 2-line almanac consists of a validity flag (*V*), an almanac number, a reduced-precision satellite clock phase offset, a set of orbital elements, and a satellite channel number. The set of orbital elements is represented as follows (all angular quantities are in semicircles and times are in seconds):- (*i*) equator-crossing longitude; inclination (offset 0.35); eccentricity; argument of the perigee; equator-crossing time; period *P* (offset from 12 h); and rate of change of period. The equator-crossing time of the reference orbit is always the first of the day. The reference day number itself occurs at the start of line 5. Day 1 corresponds to the first day of the year at the start of the 4-year leap cycle (currently 1 January 1992). The reference day number does not necessarily change each day—a set of almanacs is often allowed to remain unchanged for 2 days. The almanac number just referred to ranges from 1–25, each number in sequence (except the last) referring to a satellite location within one of the three reference planes. The first 8 almanac numbers refer to location within the orbital arc in plane one starting with phase "0" and working around clockwise in steps of 45 deg. The

Table 3 Satellite almanacs

NAVSTAR	GLONASS
Week of validity	Day of validity
Identifier	Channel number
Eccentricity	Eccentricity
Inclination	Inclination
Time of almanac	Equator time
Health	Validity of almanac
Right ascension (RA)	Equator longitude
Rate of change of RA	——
Root of semimajor axis	Orbital period
Argument of perigee	Argument of perigee
Mean anomaly	——
——	Luni–solar term
Time offset	Time offset
Frequency offset	——

second set of 8 almanac numbers (9–16) refers to satellites in plane two, and the third set of 8 almanac numbers (17–24) to satellites in plane three.

Formally, we would expect the primary purpose of almanac data to be to allow the user to predict in fairly crude terms which satellites are above his local horizon at a given time and whether their geometry is favorable for navigation. This end would be achieved by almanacs giving a position of each satellite to within 100–200 m and is the case with the NAVSTAR almanac. However, the inclusion in the GLONASS almanac of a luni–solar correction term implies a position error perhaps an order of magnitude better than a NAVSTAR almanac. Almanacs are valid for several days in the case of NAVSTAR; they are usually, but not always changed, every day in GLONASS at local midnight. It is interesting to observe that the GLONASS almanacs differ from the earlier Cicada almanacs in one major respect—the earlier almanacs were based on an equinoctial Kepler set where eccentricity and argument of the perigee are transmitted as $h = e \sin \omega$ and $l = e \cos \omega$. The equinoctial set of elements is suitable for orbits with small eccentricity, because it leads to equations with no singularities when e tends to zero.

H. GPS/GLONASS Onboard Clocks

Both GPS and GLONASS will offer, independently, precise location and time transfer continuously anywhere in the world and, indeed, in space itself. Many potential users, in particular the civil aviation community, are keenly interested in a joint GPS/GLONASS operation, because it would offer substantial advantages in defining and maintaining the integrity of the navigation aid. The question arises of compatibility of GPS/GLONASS from the point of view of satellite onboard clocks, their system references, their national standards, and, ultimately, UTC. GLONASS provides worldwide time dissemination and time transfer services in the same manner as NAVSTAR GPS, with both exhibiting substantial advantages over other existing timing services. Time transfer is both efficient and economic in the sense that direct clock comparisons can be achieved via GLONASS between widely separated sites without the use of portable clocks. Event time tagging can be achieved with the minimum of effort, and users can reacquire GLONASS time at any instant because of the continuous nature of time aboard the satellites.

Time transfer from GPS/GLONASS is achieved in a straightforward manner. Each satellite transmits signals referenced to its own onboard clock. The GPS Control Segment monitors the satellite clocks and determines their offsets from the common GPS/GLONASS system time. The clock offsets are then uploaded to satellites as part of their transmitted data message. A user at a known location receives signals from a satellite and by decoding the datastream modulated on to the transmission, is able to obtain the position of the satellite, as well as the satellite's clock offset from the common system time. Hence, the signal propagation time can be calculated at any instant. The time at which the signals are transmitted is also contained in the data message; by combining this with the propagation time and correcting first for atmospheric effects and other delays, and then for the satellite's own clock offset, the user can effect transfer to GPS/GLONASS system time. Correction to an external time scale [such as

UTC(USNO) or UTC(SU)] is then possible because the relevant offset is one of the transmitted data parameters. Any other user who has the same satellite visible is also able to transfer to the same common time scale. Clearly, if two users access the same satellite simultaneously (known as common-view reception), the difference between the two users' measurements eliminates the systematic errors common to both, such as satellite ephemeris error. In this way, time transfer between users in common view offers increased accuracy. In fact, global networks of GPS stations currently exist for the purpose of comparative ranging For example, time transfer using GPS satellites in common view is organized according to an international, global schedule by the Bureau International des Poids et Mésures (BIPM) acting as the coordinating center in Paris. A similar schedule for GLONASS is under preparation; when ready, the capability of transferring international time standards via GPS or GLONASS will result in improved measures of international time. In turn, this improved coordination of time will lead to improvements in our capability to fix position using GNSS.

In an attempt to compare GPS and GLONASS system time references, a series of measurements was conducted of the difference between each reference and UTC(USNO) using a prototype single-channel GLONASS/NAVSTAR GPS receiver. Time comparisons between system times are referenced to a 1–pulse-per-second (1 pps) strobe synchronized to UTC(USNO). The NAVSTAR system time to UTC(USNO) comparison is used as a calibration of the measurement, because the offset between GPS time and UTC(USNO) is already known—it is transmitted as part of the GPS data message.

Table 4 shows a set of measurements over a typical 24-h period on 26 October 1990. Each individual measurement lasts 180 s; satellites are accessed many times in the course of the day. The data have been corrected for tropospheric,

Table 4 NAVSTAR GPS and GLONASS system time offset from UTC(USNO)

Date	Satellite	Readings, 1-s	Average offset/ns	Standard deviation/ns
26/10/90	NAVSTAR 2	2336	−194	76
26/10/90	NAVSTAR 3	2157	−195	49
26/10/90	NAVSTAR 6	3413	−242	54
26/10/90	NAVSTAR 9	4850	−228	72
26/10/90	NAVSTAR 11	4312	−185	55
26/10/90	NAVSTAR 12	3058	−214	51
26/10/90	NAVSTAR 13	4300	−192	58
26/10/90	NAVSTAR 19	538	−205	50
26/10/90	GLONASS 34	3224	5425	58
26/10/90	GLONASS 36	4291	5444	64
26/10/90	GLONASS 39	3052	5449	60
26/10/90	GLONASS 40	4111	5437	71
26/10/90	GLONASS 41	4484	5436	54
26/10/90	GLONASS 44	4130	5478	65
26/10/90	GLONASS 45	3585	5436	62
26/10/90	GLONASS 46	4828	5437	55

relativistic, and Earth rotation effects but not for ionospheric effects. Only two of the available GPS Block II satellites (NAVSTAR 2 & 19) were used; the absence of SA on both at this time is noticeable. Both sets of data are consistent in the sense that all eight satellites individually produce results that differ from the average by much less than the standard deviation. This validation is an important feature of the measurement, because most of the measurement equipment is common to both GPS and GLONASS. By means of the transmitted offsets *A*0, it is possible to deduce a value for UTC(USNO)–UTC(SU) obtained by the satellite navigation systems GPS and GLONASS with an uncertainty of less than 50 ns. Current research is aimed at reducing the uncertainty in these measurements to the order of 10 ns.

It is known that GPS satellites carry two cesium and two rubidium atomic clocks as frequency/time standards. GLONASS satellites carry three cesium standards. The question arises as to the comparative performance of onboard GPS and GLONASS clocks. Data on the performance of certain GLONASS satellites are available.[7] Over the years 1986–1989, a steady improvement in performance has been demonstrated with clocks on board spacecraft launched during 1989 showing the qualities of high-quality cesium standards of roughly the same level of performance as the GPS block I cesiums. Since 1989, the level of performance of onboard clocks has been consistently high. It is planned to use improved cesium clocks on future GLONASS satellites. These new clocks, known as "Malachite" atomic standards will offer long-term stabilities five times better than those currently in operation.

In determining the accuracy with which time (or its equivalent, pseudorange) can be measured, it is important to remember that two basic quantities are available: 1) code phase; and 2) carrier phase. In crude terms, we can measure code phase, meaning the time interval between local and transmitted code epochs, to an accuracy limited by the code frequency, the signal-to-noise ratio, and the bandwidth of the measurement. Ultimately, if all satellite onboard frequency sources are locked to the same frequency standard, the accuracy available is limited by the stability of that standard. In the case where onboard standards are synchronized atomic clocks, range, and hence, position can be established to a fraction of a wavelength at the carrier frequency. At *L*-band frequencies of 1.6 GHz, the free-space wavelength is around 19 cm. Because carrier-phase measurements are fractional, range to a satellite can only be found by inclusion of the integer number of carrier wavelengths between the satellite and observer. In practice, the integers may only be resolved if two or more receivers are operating simultaneously (differential operation) on the same satellite.

It is also important to keep in mind the rate at which ranging measurements can be made. Both GPS and GLONASS transmit a timing epoch every millisecond. Taken in conjunction with the fact that most modern receivers are designed as multichannel units, this implies an ultimate receiver capability of producing raw pseudoranges to all visible satellites simultaneously 1000 times a second. In most applications, such a high rate of raw measurement is not necessary. However, the point does demonstrate that the full capability of GNSS is far from being reached.

II. Performance of GLONASS and GPS + GLONASS

A. Introduction

In this section, we examine the performance capabilities of GLONASS and compare them with those of GPS. The GPS performance has been discussed fully in the companion volume, Part II. We include some additional results here for a side-by-side comparison with GLONASS, where appropriate. The context for our performance analysis is civil aviation. The interest in GLONASS stems from the recognition that GPS alone falls short of meeting the requirements of a global sole-means, or stand-alone, navigation system, as discussed below; GLONASS alone does the same. The two systems taken together, however, offer amply redundant measurements to all users, and seem capable of meeting these requirements globally. Using the combined set of measurements from GPS and GLONASS for positioning is referred to as an integrated use of the two systems, and denoted as GPS + GLONASS.

A user equipped with a GPS + GLONASS receiver may consider the two systems as augmenting each other. Other approaches to augmentation of either system are based on LORAN, inertial reference system, and baroaltimeter, as discussed in Chapters 6–8, this volume. Each approach brings about an improvement in the availability of a GNSS-based navigation service. Currently, the most important of these augmentations appears to be the FAA's Wide Area Augmentation System (WAAS), which is scheduled to provide a sole-means capability over the conterminous U.S. starting in 1998 (Chapter 4, this volume). If successful, WAAS would to expand to other countries and is expected to evolve into a seamless global augmentation of GPS.

The performance measures relevant to positioning are: distribution of the number of satellites in view and a characterization of their geometry; and the quality of the measurements. The performance capabilities of GLONASS, are substantially similar to those of GPS. The main difference between the two is SA. GLONASS, which has disavowed an SA-like feature, offers a significantly better positioning accuracy than does GPS with SA active. On the other hand, at this writing GPS is close to achieving operational-status, while the prospects of GLONASS are less clear.

We begin with a brief review of the requirements of a civil aviation navigation system, followed by a discussion of the technical issues related to an integrated use of the two autonomous systems. In Sec. II.C, we discuss the performance achievable from GLONASS and from integrated use of GPS and GLONASS via-à-vis the requirements of civil aviation. We focus exclusively on the signals planned to be available for civil use; namely, C/A-code on L_1. The data analysis results are based on nearly continuous measurements from the two systems since 1990 and have been reported in greater detail elsewhere.[8,9] While GPS receivers have been widely available, GLONASS receivers have remained rare. MIT's Lincoln Laboratory has had one or more of the following GLONASS receivers since 1990: a prototype receiver built by the Magnavox Corporation; ASR-16, an aviation model from the erstwhile Leningrad Radiotechnical Research Institute; and R-100, a GPS + GLONASS receiver under development at 3S Navigation.

B. Requirements of Civil Aviation

For a navigation system to be adopted for use in civil aviation, it must meet certain stringent criteria. The criteria are stated as standards and certification procedures for each piece of equipment installed in the cockpit or deployed at the airports or elsewhere for use in navigation. International civil aviation requires agreement on these standards and procedures among the national and regional regulatory agencies.

The requirements basically relate to three areas: coverage, accuracy, and integrity monitoring. The coverage of a navigation system deals with where and when can the system be used. A satellite-based navigation system is usable for three-dimensional positioning when four or more satellites are in view of the user. A global system must, therefore, deploy a large enough constellation so that all users at different locations and times see enough satellites. The requirement on accuracy refers to the positioning accuracy provided by the system. The accuracy requirements in civil aviation depend upon the phase of the flight, and they currently range from several kilometers for en route phase to several hundreds of meters for a nonprecision approach. The precision approaches, executed under poor visibility conditions, require that the navigation system guide an aircraft down to an altitude of 60 m or less. These approaches, currently executed on specially equipped runways, require much more precise position estimates. Although satellite-based navigation seems promising for precision approaches also, these are outside the scope of this chapter.

The requirement on integrity monitoring deals with an issue vital to civil aviation: the ability of the system and its users to detect a system malfunction in a timely manner. The main point is that the user must be able to rely on the position estimate provided by the system. A system may be certified as *supplemental* or *sole means.* A supplemental system must provide a position estimate of the required accuracy, when it can, and recognize a situation when it cannot. In the latter case, the system must warn the user, who may then switch to an alternate system for navigation. A sole-means system, as the name suggests, should require no other navigation system as a backup. Therefore, a sole-means system, or its users, must be able to recover from the possible malfunctions. Obviously, the idea of a sole-means system is economically attractive, and the integrated use of GPS and GLONASS is seen as a potential sole-means system. Indeed, if this promise could be shown to be met, there would be no need for any of the current ground-based navigation aids: VHF omnidirectional radio (VOR), distance-measuring equipment (DME), LORAN, and OMEGA. This was seen as particularly important because at present there are no ground-based navigation aids over long stretches in economically underdeveloped parts of the world and in sparsely populated areas (e. g., Alaska and parts of Russia and Canada).

Integrity monitoring is discussed in detail in Chapter 5, this volume. For completeness, the main ideas are reviewed below. The integrity-monitoring requirements for a satellite navigation system are typically stated as follows. If the error in a position estimate exceeds a certain threshold, the user must be notified within a certain time interval. Both the error threshold and the required response time depend upon the phase of the flight and can range widely. The system failure scenario for a satellite navigation system is defined as an erroneous or out-of-tolerance signal transmitted by one of the satellites in the constellation. The current view is that the constellations would be managed so that at any instant the probability of two or more satellites transmitting anomalous signals simultaneously while marked as healthy is considered negligible. Note, however, that even if a system is known to be operating to specifications, a user has to guard against an unacceptable position estimate obtained because of poor geometry or poor measurement quality.

The current accuracy and integrity-monitoring requirements for the various phases of flight are discussed in Chapters 12–14, this volume. For the purpose of evaluation of GPS and GLONASS performance, recall that: 1) in en route and terminal phases of flight and during a nonprecision approach, a navigation system is required to provide only a two-dimensional location of the aircraft; altitude is provided by a baroaltimeter; and 2) for nonprecision approach, the alarm limit for the horizontal error is 0.3 n.mi. (555 m), and time to alarm, 10 s.

C. Integrated Use of GPS and GLONASS

GPS and GLONASS are autonomous systems, each with its own time scale and coordinate frame in which to express a three-dimensional position. As discussed in Sec. I, the time scales adopted by the two systems are their national standards: UTC(USNO) and UTC(SU). The offset between the two time scales has been stable in recent years, but this stability cannot be taken for granted.

Clearly, a user interested in the integrated use of the two systems must be able to determine the instantaneous difference between the two time scales. The problem can be thought of as one of position estimation from two sets of pseudoranges, each with an unknown time bias, which makes five unknowns in all. Obviously, one or both systems could carry information on this bias as a part of their navigation messages. At worst, without the information on bias, we could solve for the additional unknown by "sacrificing" a measurement. As we shall see, the integrated use of GPS and GLONASS offers amply redundant measurements, and the additional unknown does not create a problem.

The two systems express the positions of their satellites, and, therefore, of their users, in different geocentric coordinate frames. In 1987, GPS adopted the WGS84 system.[10] GLONASS started with the SGS85 system[11] but switched in 1993 to PE-90. Our current knowledge of both SGS85 and PE-90 is quite limited; the differences between the two are believed to be insignificant for civil aviation purposes.

Combining measurements from GPS and GLONASS requires that we estimate a transformation between their coordinate frames. Estimation of the transformation is straightforward in principle; it requires positions of a set of points expressed in both coordinate frames. While a point on Earth can now be surveyed to centimeter-level accuracy in WGS84 using GPS measurements, the corresponding SGS85/PE-90 coordinates have been difficult to determine. The main difficulty has been the lack of precise and sturdy GLONASS receivers.

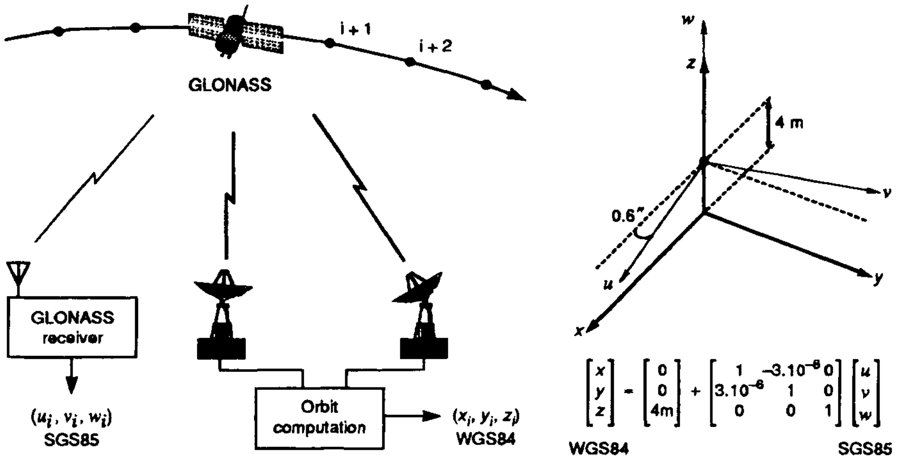
Access to GLONASS receivers and the facilities of the Deep-Space Tracking Network (DSTN) operated by MIT's Lincoln Laboratory gave us the resources to take a different approach to this problem in 1992. We took advantage of the fact that the positions of GLONASS satellites as defined in SGS85 were available to us as a part of the navigation messages broadcast by the satellites. The remaining task, then, was to obtain the corresponding coordinates in WGS84, and that is where the resources of the DSTN came into play. In experiments, we tracked several GLONASS satellites independently to characterize their ephemerides in WGS84 and compared these to the satellite positions in SGS85 as broadcast by the satellites themselves and recorded by the GLONASS receivers. The results showed that the coordinates of points on Earth as expressed in the two coordinate frames differ by no more than 20 m,[12] and that the two geocentric coordinate frames are brought substantially into coincidence by a small rotation (0.6″) of the z-axis, and a small displacement of the origin. Figure 4 illustrates the process of gathering the position data in the two coordinate frames, and the resulting estimated transformation.

With the time and space reference standards reconciled, the design of a receiver to obtain measurements from both GPS and GLONASS poses no basic challenge. That such receivers remain rare is attributable mainly to the present uncertainty about the future of GLONASS.

D. Performance of GLONASS and GPS and GLONASS

We discuss next the level of performance achievable from GLONASS and from an integrated use of GPS and GLONASS. In particular, we review coverage, accuracy, and integrity-monitoring capabilities and compare the performance in each of these areas with the requirements of civil aviation.

There is uncertainty at present about the size and management of the GLONASS constellation. GLONASS was defined initially as a constellation of 21 active satellites plus up to three on-orbit spares. More recently, however, it seems that both GLONASS and GPS may be planning to maintain a 24-satellite constellation.[13] We present performance results for an average 21-satellite GLO-


i + 1
i + 2
GLONASS
GLONASS receiver
(u_i, v_i, w_i)
SGS85
Orbit computation
(x_i, y_i, z_i)
WGS84
w
z
4 m
v
0.6″
u
y
x
$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 4\text{m} \end{bmatrix} + \begin{bmatrix} 1 & -3.10^{-6} & 0 \\ 3.10^{-6} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}$
WGS84
SGS85

a) Data collection

b) Coordinate transformation

Fig. 4 Position data-gathering process and resulting estimated transformation.

NASS constellation (GLONASS-21) and for the average GPS and GLONASS constellations combined: GPS + GLONASS (2 × 21). These results are obtained in simulations with three randomly chosen satellites in each constellation declared as unavailable in each trial.

1. Coverage

Figure 5a shows the distribution of the number of satellites visible to a user at a random location on Earth at a random time, counting only those satellites that are well above the horizon (i.e., elevation > 7.5 deg).

The coverage results for GLONASS-21 are quite similar to those for GPS-21. About 0.5% of the users would see fewer than four satellites; nearly 20% fewer than six; and about 50% fewer than seven. With the combined constellations, however, all users see 7 or more satellites, 99% see 10 or more, and nearly half see 14 or more! Clearly, some users may not be able to estimate their positions using GLONASS (or, GPS) alone. With the combined constellation, however, *all* users will have abundantly redundant measurement sets on which to base a position estimate.

Figure 5b gives the cumulative distribution functions (cdf) of the horizontal and vertical dilutions of precision (HDOP and VDOP, respectively) available globally from GLONASS-21 and GPS + GLONASS (2 × 21). As discussed in Chapters 5 and 9, the companion volume, DOP parameters characterize the quality of the position estimates available from a constellation of satellites in view. The cdf in Fig. 5b define the availability of favorable satellite geometries for position estimation, and are to be interpreted in view of the relationship: rms position error = DOP. σ_{URE}, where σ_{URE} is the rms value of the user range error (URE). We discuss this relationship further in the next section, and indeed, use it to estimate σ_{URE} from position error data.

The distribution of VDOP has been included in Fig. 5b for completeness; we concern ourselves mainly with HDOP and horizontal accuracy for the reason noted earlier. The distributions of HDOP and VDOP for GLONASS are similar to those for GPS. With GLONASS-21, satellite geometries characterized by HDOP < 2 would be available to nearly 94% of the users; with GPS + GLONASS, such favorable geometries would be available to *every* user. Civil aviation, however, has no requirements related to DOPs, only to position accuracy, to which DOP is a contributor. The other contributor is measurement data quality, which we examine in the next section.

2. Accuracy

We turn next to the accuracy of the position estimates obtained from GLONASS and examine the quality of the position estimates based on measurements recorded in our laboratory. Figure 6a is typical of GLONASS. For comparison, we have also included a corresponding display of the quality of position estimates obtained from GPS (Fig. 6b). Figures 6a and b were both generated in the same way. Snapshots were taken at 1-min intervals of range measurements from the satellites in our view from each constellation over a period of 24 h on 31 December 1993 and position estimates computed when possible. The discrepancy in each position estimate was computed relative to the known, surveyed location of the

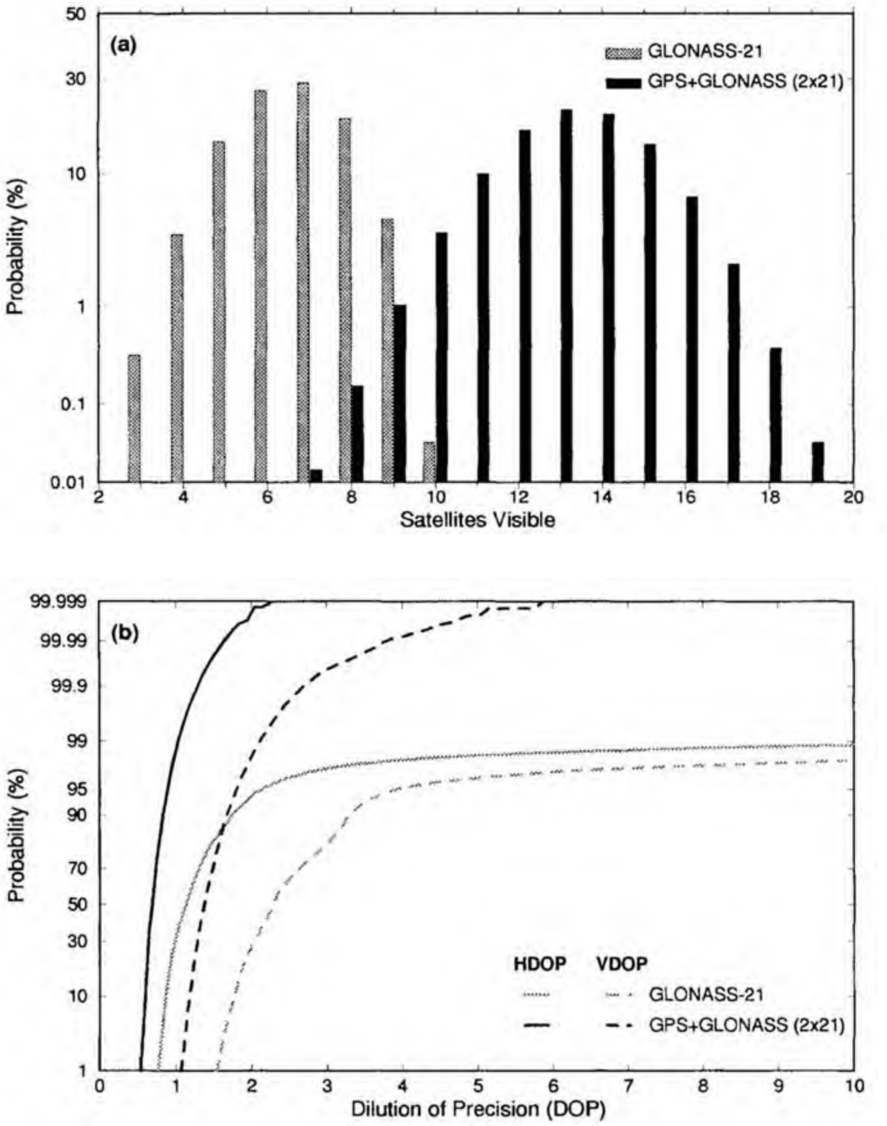


Fig. 5 Coverage provided by GLONASS and GPS + GLONASS.

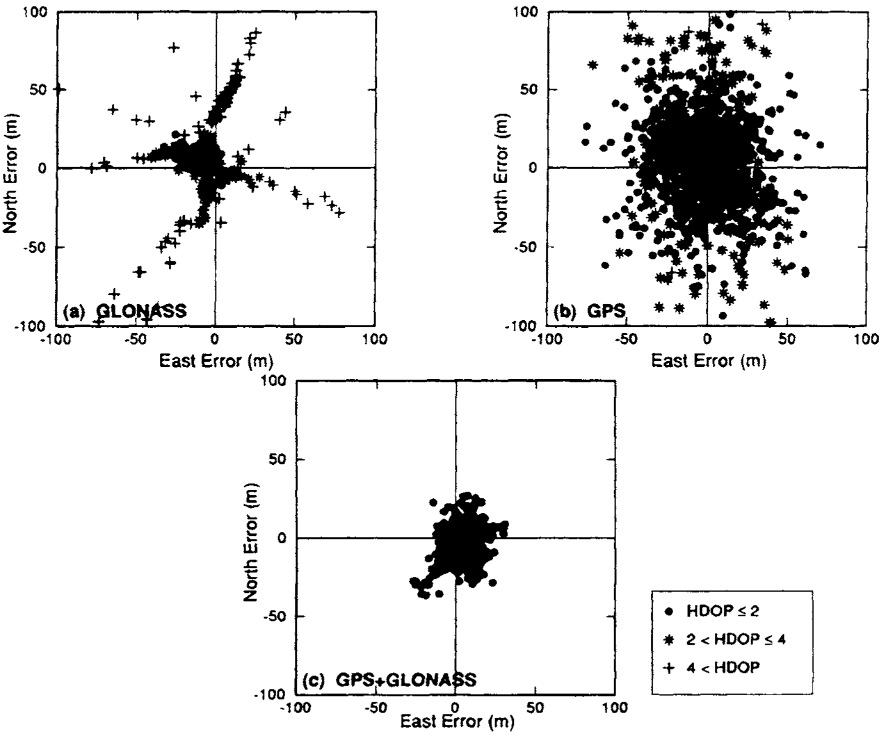


Fig. 6 Position estimates from GPS, GLONASS, and GPS + GLONASS (1-min samples, 31 December 1993).

antenna in the WGS84 coordinate frame, and the horizontal components of the discrepancy were plotted. Because the accuracy of a position estimate depends upon the satellite geometry at the time, each point is coded to reflect the corresponding HDOP. On 31 December 1993, GPS had 24 satellites in orbit, 20 of which were subject to SA; GLONASS had 12 satellites marked healthy. With measurements limited to satellites with SA active, position estimates could be computed from GPS nearly 90% of the time. The leaner GLONASS constellation provided the position estimates about half the time.

Figure 6a shows a tight cluster of position estimates corresponding to favorable satellite geometries (HDOP < 2). As noted earlier, such geometries would be available globally to 94% of the users with an average 21-satellite GLONASS constellation. The straggling position estimates shown correspond to poor satellite geometries (HDOP > 4). With the current sparse GLONASS constellation, the proportion of these points is relatively large, but with the operational constellation, fewer than 2% of the users would encounter such geometries. That the tight cluster in Fig. 6a is off-center was expected, because of the difference in the coordinate frames referred to earlier. The observed difference is consistent with our estimated transformation between SGS85 and WGS84.

In Fig. 6b, with SA on, the GPS position estimates appear to be widely scattered, as compared to Fig. 6a. These position estimates, however, are consistent with the GPS specifications on horizontal accuracy available to civil users. The strength of the constellation is reflected in the relatively small proportion of the symbols corresponding to HDOP > 2. Actually, HDOP was below 2 for nearly 90% of the measurement samples.

Finally, we look at the position estimates obtained from the combined set of measurements from GPS and GLONASS. The results, presented in Fig. 6c, illustrate the reason for our interest in GLONASS. Figure 6c combines the best features of Fig. 6a and b, and is indeed a distinct improvement over both. GPS contributes a larger satellite constellation, and GLONASS, measurements of better quality. The result is consistently good satellite geometries and mitigation of SA. Of course, with a full GLONASS constellation, the results would be better yet.

Figures 6a–6c are intended as a qualitative view of the positioning capabilities of GLONASS and GPS. To be able to predict the positioning accuracies achievable from these systems, when operational, we must take a closer look at the measurement error in the two systems. As noted earlier, we use the well-known relationship: rms position error = DOP. σ_{URE}, to estimate the rms value of the range measurement error for each system. From each snapshot of the measurements from a surveyed antenna location, we compute a DOP value for the measurement set and the error in the corresponding position estimate. Figures 7a and 7b give plots of horizontal error vs HDOP from measurement snapshots taken 3 minutes apart over a period of about three months in 1993. These scatter plots illuminate the nature of the often misunderstood relationship between DOP and position error. We divide the points in HDOP bins and compute the rms value of the position error in each bin. It is no surprise that the relationship between the rms position error and DOP is linear with slope σ_{URE}.

As seen in Figure 7a, the rms URE for GLONASS is approximately 10 m, and it has remained substantially at this level over the past 3 years. With SA on, the URE for GPS is apparently changeable, but it has been relatively stable during 1992–1993, with an rms value of 25–27 m. With SA off, σ_{URE} for GPS has remained at about 7 m. We have also analyzed the range measurements from the two systems over an extended period to characterize the cdf of the range error for each.[14] The difference in σ_{URE} between GLONASS and GPS (SA off) is attributed mainly to the fact that GPS transmits in its navigation message the values of parameters to compensate partially for the ionospheric delays on the basis of a model; GLONASS does not.

We now have all the elements necessary for a global view of the positioning accuracies achievable from the operational GPS and GLONASS systems, separately and together. To recapitulate, the error in a position estimate is determined by the spatial distribution of the satellites around the user (satellite geometry) and the error in the range measurements, and we now have a complete characterization of both. Table 5 summarizes the global projections for the quality of the position estimates available from "average" 21-satellite constellations of GPS and GLONASS, when operational, on the basis of their performance as observed over 1992–1993.

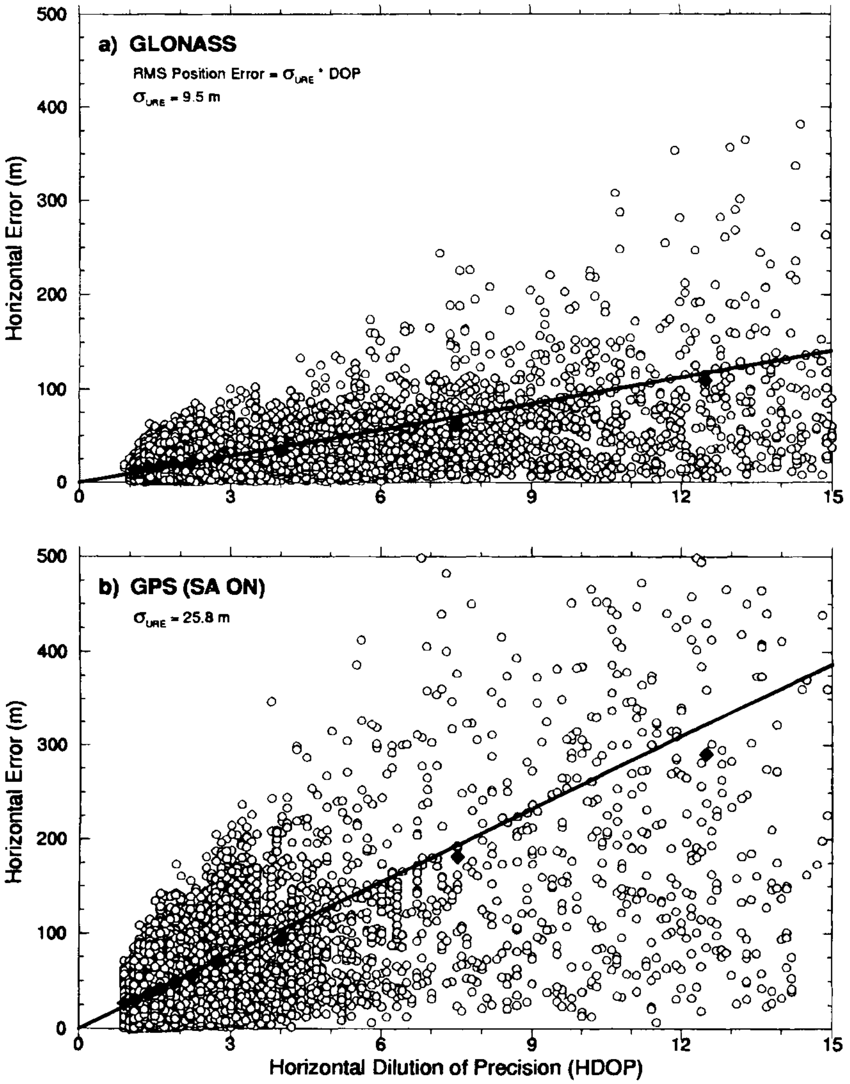


Fig. 7 Estimation of the rms value of the User Range Error (URE) from measurements taken over a 3-month period in 1993.

Table 5 **The projected positioning accuracy of GPS and GLONASS**

	Horizontal error		
	50%	95%	Vertical error, 95%
GPS (SA off)	7 m	18 m	34 m
GPS (SA on)	27	72	135
GLONASS	10	26	45
GPS + GLONASS	9	20	38

3. *Integrity Monitoring*

The navigation accuracy results of Table 5 assume that each system is operating to specifications. A user, however, cannot take this premise for granted. Indeed, both GPS and GLONASS have an extensive self-diagnosis capability on board the satellites, as well as monitoring facilities at the ground control stations. It is unclear, however, if an error can be detected and the appropriate flags set in the navigation message transmissions quickly enough to suit a pilot who is using the satellite signals in preparation for a landing. Basically, a critical demand of civil aviation is that the navigation system provide not only a position estimate but also an assurance that the estimate is "good" (i.e., the position error does not exceed a tolerable level). The idea of guarding against anomalous position estimates is called system integrity monitoring.

An approach to integrity monitoring of a satellite navigation system is to base an inference on the accuracy of a position estimate on the measurements themselves. The idea is to verify that the measurements are consistent with a model and to characterize the quality of a position estimate. This can be done if the measurement set is redundant (i.e., we have more measurements than the minimum needed for position estimation.) An important benefit of this approach, which is known as *receiver autonomous integrity monitoring* (RAIM), is that it eliminates the need for an external means of detecting system malfunction and a communication network to disseminate this information to the users.

The problem of detection and isolation of an anomalous range measurement may be thought of as one of detecting inconsistency in a set of linear equations and then identifying the anomalous equation. Obviously, at least one redundant equation is required to detect the presence of an anomaly via a consistency check. Similarly, at least two redundant equations are required to identify the anomalous equation. These, however, are only the necessary conditions, and do not guarantee an effective consistency check. The effectiveness of the check depends upon the conditioning of the set of equations and their subsets. Our task is complicated further because the equations are only approximate, being based on range measurements that include errors, the sources of which have been cited earlier. As an aside, note that the DOP parameter basically reflects the notion of linear independence of the direction vectors to the satellites and conditioning of a set of measurement equations.

Because at least four satellites are required to be in view to compute a three-dimensional position, users with five or more satellites in view may be able to use the system as supplemental. On the other hand, the real economic payoff will follow the adoption of a satellite navigation system as sole means. However, this would require that all users have six or more satellites in view. Based on the satellite visibility results for GPS and GLONASS (Fig. 5), each system by itself falls considerably short as a candidate for a sole-means system. Both systems taken together, however, offer amply redundant measurements and potential for RAIM and a sole-means system.

According to the integrity requirements, a supplemental navigation system must provide each user with a position estimate of the required accuracy, or an indication that it cannot. Obviously, the more often a system is usable, the better. If it is usable 100% of the time, we have a sole-means system. By definition, the users of a sole-means system must be capable of recovering from system failure.

A simple RAIM scheme could work as follows. Figure 8 is an azimuth-elevation sky map of the satellites, shown as stars, from the two constellations in view of a user. The total number of measurements available to a user is considerably larger than the minimum required. This is consistent with our previous results (Fig. 5). Suppose that the satellite shown as a dark star in Fig. 8 is providing anomalous measurements undetected by the system and unknown to the user. A user tracking the eight satellites inside the solid ellipse could compute a position estimate by using all of the measurements and, as a check on its quality, could compare it with the eight additional position estimates obtained when leaving out one measurement at a time. As noted earlier, the quality of a position estimate depends upon two factors: the error in the range measurements and the geometry of the satellites. Although unknown to the user, all eight satellites are performing to specifications, and the satellite geometries involved in all nine position estimates are uniformly good. The nine position estimates form a tight cluster, as shown on the right side of Fig. 8. This reassures the user of the consistency of the measurements and the quality of the position estimate.

On the other hand, suppose a user tracking the eight satellites inside the dashed ellipse in Fig. 8, including the faulty satellite, were to try this same check. The

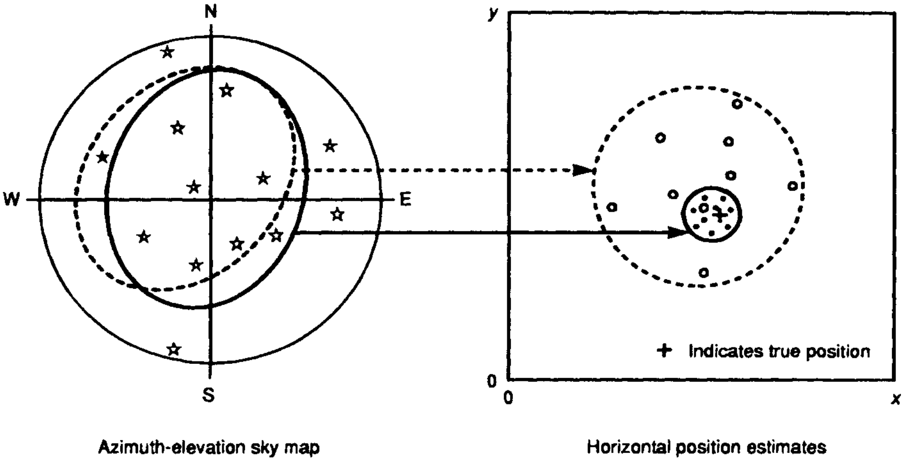

N
W
E
S
y
0
0
x
+ Indicates true position
Azimuth-elevation sky map
Horizontal position estimates

Fig. 8 Receiver autonomous integrity monitoring on the basis of scatter in the position estimates obtained from subsets of measurements.

resulting cluster would be larger; the actual size would depend primarily upon the size of the error in the faulty measurement. A user, if assured of good satellite geometries associated with the position estimates computed as a part of this check, could treat the size of the cluster as a predictor of the quality of the position estimate.

We have pursued this approach to RAIM, and have developed an algorithm for a position estimate and a measure of its quality, given the probable failure scenario discussed earlier. We define the measure of quality as a high-confidence estimate of an upper bound on the error in the position estimate and call it the *integrity level.* The integrity level is defined as $P\{\text{position error} < \text{integrity level}\}$ $<\epsilon$, where ϵ is a suitably low, user-defined parameter. To be usable, the integrity level must be a "tight" error bound consistent with the required alarm limits. Two other essential points must be mentioned. First, the above relation is to be interpreted as a conditional probability, given that one of the satellite measurements could be in error by an indeterminate amount. The total probability that a position error could exceed its associated integrity level would be even lower; namely, it would be the probability of a system failure (expected to be quite rare) multiplied by ϵ. Second, the ability to compute the integrity level is predicated on the availability of measurements from n satellites (where $n \geq 5$), assuring good geometries for each subset of $(n - 1)$. To obtain a "tight" integrity level, we require at least five satellites satisfying the above requirements on geometry and operating to specifications.

Figure 9 illustrates the idea to be implemented. Given a set of measurements, the user computes a position estimate and its associated integrity level. The estimate is acceptable if the integrity level does not exceed the alarm limit for that phase of flight. If a system can assure all its users of the integrity levels they require at all times, then we have a sole-means system. The RAIM algorithm, described in greater detail in Ref. 14, consists of the following steps. Select n

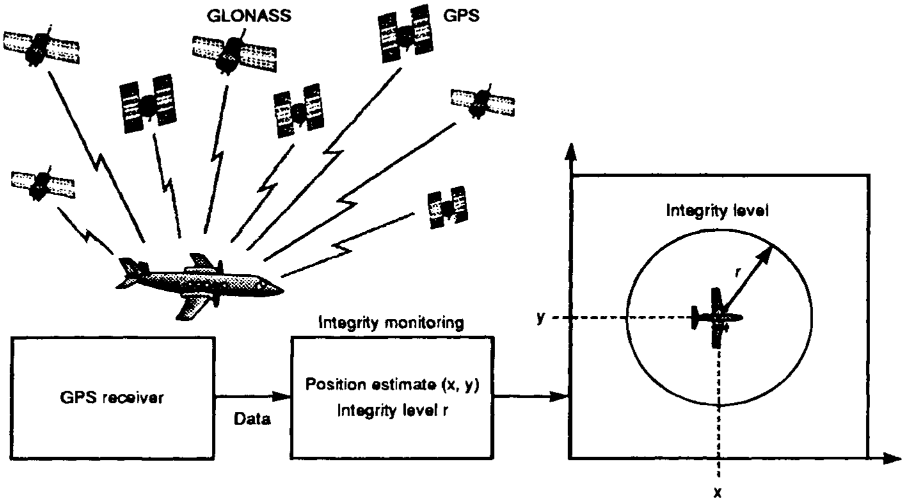


Fig. 9 Computing a position estimate and its associated integrity level.

satellites (where $n \geq 5$) among those visible, estimate positions from all n measurements, and from $(n - 1)$ measurements at a time, determine the size of the cluster (i.e., our RAIM statistic) formed by these position estimates, and obtain the corresponding integrity level from a precomputed table. If the integrity level is unsatisfactory to the user, then switch satellites for a better estimate, if possible. The computation of the table of scatter vs integrity level is at the heart of the algorithm. It requires computation of the conditional probability distribution function of the position error computed from a measurement set containing a faulty measurement, given the scatter.

Figure 10 gives the distribution of integrity levels available to the users of GPS + GLONASS worldwide corresponding to the value of 10^{-5} for ϵ. The conclusion evident in this figure is that the combined measurements from the GPS and GLONASS systems offer a comfortable level of redundancy so that even if one of the measurements is anomalous, 99.9% of the users would be able to compute position estimates with an assurance that their position error does not exceed 200 m. Nearly all users would be able to obtain position estimates with an error below 500 m, meeting the requirements for a nonprecision approach. If SA were to be switched off in GPS, the estimates would be significantly better, as also shown in Fig. 10. This performance corresponds to the 2×21 constellation of GPS + GLONASS and reflects other assumptions on measurement quality and constellation availability that are believed to be on the safe side. By using a RAIM-based approach, therefore, GPS + GLONASS is expected to meet the requirements of a sole-means navigation system for en route and terminal phases of flight, and for nonprecision approaches.

We should note that although GLONASS, like GPS, falls short of meeting the requirements of a sole-means system, it can be used as a supplemental navigation system. GLONASS-21 can offer nearly 90% of the users the integrity level required for nonprecision approaches; GLONASS-24, 99%.

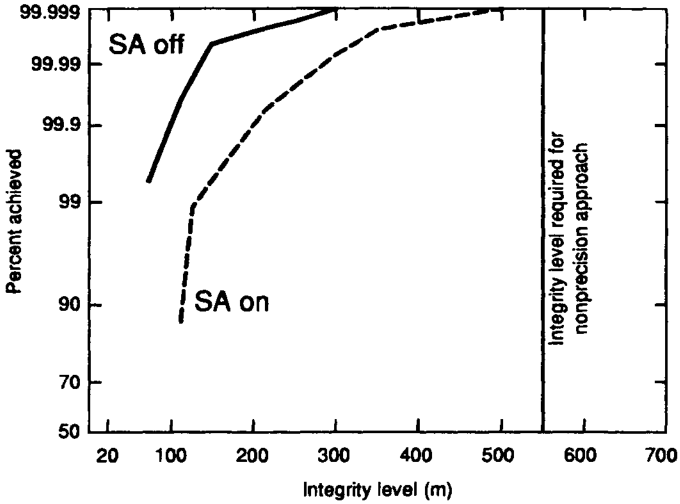


Fig. 10 Distribution of integrity levels available to the users of GPS + GLONASS.

III. Summary

The Global Navigation Satellite Systems GPS and GLONASS have much in common in terms of their signal designs, orbital plans, and ownership by military. A principal difference apparent in December 1994 is that GPS is very close to achieving operational status; GLONASS is planned by the Russian Federation to follow, but the development schedule remains unclear. This uncertainty is reflected in the current lack of ready availability of commercial GLONASS receivers anywhere, including Russia.

The performance of GLONASS in terms of coverage and positioning accuracy is potentially similar to that of GPS in the absence of SA. Like GPS, GLONASS is an attractive candidate as a supplemental navigation system for civil aviation. The combination of signals from the two systems, however, is a potential sole-means system for en route and terminal phases of flight, and for nonprecision approaches globally. The performance achievable in integrated use in coverage, accuracy, and integrity monitoring seems capable of meeting the requirements of a sole-means satellite navigation system.

While GLONASS deployment has lagged, the Wide Area Augmentation System (WAAS) has emerged as an attractive approach to achieving a sole-means capability over large geographic areas. The acquisition of WAAS is underway in the US, and several other countries have expressed an interest in expanding it beyond the conterminous United States (CONUS). This development notwithstanding, GLONASS remains potentially an attractive autonomous system of great value to the civil community.

Acknowledgments

The ISN Leeds program of research on global navigation satellite systems has been supported by the UK Civil Aviation Authority (CAA). The work on GLONASS performance at MIT Lincoln Laboratory was sponsored by the Federal Aviation Administration.

References

[1] Anon., “Global Positioning System,” *Navigation,* Special Issue, Summer 1978.

[2] Anodina, T. G., “Global Positioning System GLONASS,” Special committee on future air navigation systems (FANS/4), International Civil Aviation Organization, Montreal, May 2–20, 1988.

[3] Blackwell, E. G., “Overview of differential GPS Methods,” *Navigation,* Vol. 32, No. 2, 1985, pp. 114–125.

[4] Remondi, B. W., “Kinematic and Pseudo-kinematic GPS,” *Proceedings of the ION Satellite Division's International Technical Meeting,* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 19–23, 1988, pp. 115–128.

[5] Lennen, G., “The USSR's GLONASS P-Code Determination and Initial Results,” *Proceedings of the Institute of Navigation Satellite Division's International Technical Meeting,* (Colorado Springs, Co), Institute of Navigation, Washington, DC, Sept. 1989.

[6] Dale, S. A., Daly, P., and Kitching, I. D., “Understanding Signals from GLONASS Satellites,” *International Journal of Satellite Communications,* Vol. 7, No. 1, 1989, pp. 11–22.

[7] Daly, P., Kitching, I. D., Allan, D., and Peppler, T., "Frequency and Time Stability of GPS and GLONASS Clocks," *International Journal of Satellite Communications,* Vol. 9, No. 1, 1991, pp. 11–22.

[8] Misra, P., et al., "GLONASS Data Analysis: Interim Results," *Navigation,* Vol. 39, Spring 1992, pp. 93–109.

[9] Misra, P., et al., "GLONASS Performance in 1992: A Review," *GPS World,* May 1993, pp. 28–38.

[10] Anon., "Department of Defense World Geodetic System 1984: Its Definition and Relationships with Local Geodetic Systems," Defense Mapping Agency, DMA TR 8350.2, 1987.

[11] Anon., "GLONASS Interface Control Document," GLAVCOSMOS, USSR, 1991.

[12] Misra, P., and Abbot, R. A., "SGS85–WGS84 Transformation," *Manuscripta Geodaetica,* Vol. 19, 1994, pp. 300–308.

[13] Anodina, T. G., "Status of and Prospects for the Development of the GLONASS Satellite Navigation System," FANS(II)/4-WP/47, International Civil Aviation Organization (ICAO), Montreal, 15 Sept.–1 Oct. 1993.

[14] Misra, P., et al., "Receiver Autonomous Integrity Monitoring (RAIM) of GPS and GLONASS," *Navigation,* Vol. 40, Spring 1993, pp. 87–104.

Chapter 10

Land Vehicle Navigation and Tracking

Robert L. French*
R. L. French & Associates, Fort Worth, Texas 76107

I. Application Characteristics and Markets

LAND vehicle navigation and tracking systems represent one of the largest potential applications for Global Positioning System (GPS) receivers in terms of the sheer numbers that will be required over the next 20 years. Both vehicle navigation and tracking are important subsets of intelligent vehicle highway systems (IVHS),† a major worldwide movement to improve the efficiency, safety, and environmental aspects of road traffic through the application of information, communications, positioning, and control technologies.

The ability to determine vehicle location is the most fundamental requirement of both vehicle navigation and tracking systems. However, although navigation and tracking applications share certain common requirements and characteristics, they also are also distinctly different in many ways. Accordingly, this chapter starts with a brief characterization of tracking, navigation, and IVHS to establish a frame of reference for discussing these GPS applications.

A. Commercial Vehicle Tracking

Vehicle-tracking systems are often called by other names such as automatic vehicle location-monitoring systems or position-reporting systems. The term "commercial vehicle" as used here refers to public service vehicles such as those of ambulance, fire, police, and transit departments as well as to all classes of vehicles used in business and government service. With few exceptions (e.g., those that subscribe to commercial security or stolen vehicle tracking services), vehicles in ordinary consumer use are not considered to be part of the vehicle tracking market.

*Principal.

†As this book goes to press, the term "Intelligent Vehicle Highway Systems" is generally being changed to "Intelligent Transportation Systems" (ITS) in the United States.

One of the most distinct differences between vehicle-tracking and navigation systems is that tracking systems do not necessarily require that vehicle location be known to the onboard component. Rather, vehicle location must be known at a central dispatch office or monitoring station, thus allowing tracking to be accomplished by multilateration using vehicle transmitter signals arriving at fixed receiver sites. However, it is common for vehicle location to be determined by onboard equipment and then transmitted to the dispatch office over a mobile data communication link.

Depending upon application, vehicle location may be communicated automatically at programed time or distance intervals, or in response to inquiries or "polling" by the dispatch office.

Thus, tracking systems typically incorporate, or are incorporated into, mobile data communications systems that report vehicle location and may also report sensed vehicle status information (e.g., engine temperature) and èxchange messages between the driver and the central dispatch office. Most vehicle-tracking systems are closely integrated with computer-aided dispatch (CAD), routing, and scheduling, or other forms of computerized fleet management systems using the same communication link.

A second major difference between vehicle navigation and tracking systems is the accuracy and continuity required of the positioning subsystem.[1] Automobile navigation and route guidance require continuous location information accurate to within 10–25 m in order to know always exactly which street is being traveled and which intersection is being approached. Vehicle tracking, depending upon the particular application (e.g., emergency response vehicles) may be equally or even more demanding in terms of location accuracy and continuity.

At the other extreme, accuracies of 500–1000 m may be acceptable for systems installed in intercity trucks because it is often adequate for the dispatcher to know on what major highway the truck is traveling or what city it is near. Tracking systems for many classes of commercial vehicles operating in urban areas may require only intermediate levels of accuracy (typically 100–300 m) depending upon the application.

Transit buses were among the first fleet operations to start using vehicle-tracking systems. Other early vehicle-tracking applications included police fleets. Major programs were sponsored by the Urban Mass Transit Administration (UMTA) during the 1970s to evaluate electronic signpost, radio frequency pulse trilateration, Loran-C, and other approaches.[2] In 1991, major transit fleets started turning to GPS for vehicle tracking.

During the 1980s, Loran-C became popular for tracking urban fleet vehicles and was used extensively with satellite communications in early intercity trucking applications. Other forms of satellite positioning such as Geostar's failed Radiodetermination Satellite Services (RDSS) endeavor[3] and Qualcomm's successful QASPR[4] (Qualcomm Automatic Satellite Position Reporting system) for intercity trucking also came on the scene during the 1980s. However, as the Navstar satellite constellation approached completion and GPS receiver prices started becoming more competitive in the early 1990s, many vehicle-tracking system developers turned to GPS.

Figure 1 illustrates the overall concept and principal subsystems for a representative vehicle-tracking system that uses an onboard radionavigation receiver (e.g.,

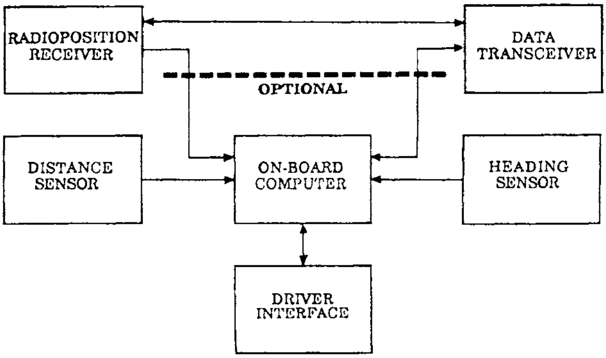


Fig. 1 Typical components and subsystems of vehicle tracking system.

GPS or Loran-C) to determine vehicle location for reporting over an arbitrary mobile data communication link. Satellite communications are typically used for tracking intercity trucking, whereas private land mobile radio or specialized mobile radio (SMR) are commonplace for urban fleet applications. (See Sec. III.E for a brief characterization of alternative mobile data communications approaches.)

Occasional discontinuities in vehicle location information, such as those encountered with GPS, Loran-C, and other radiolocation technologies in urban environments, may be a problem depending upon application, reporting protocol, and frequency of reporting. Thus, as indicated below the dashed line in Fig. 1, optional dead reckoning subsystems may be incorporated to keep track of vehicle location between valid radiolocation fixes. Although effective over short intervals, dead reckoning accuracy deteriorates with distance and/or time unless updated by GPS or other means, such as map matching in the case of automobile navigation. (See Sec. III.A for a brief description of dead reckoning approaches for vehicles.)

B. Automobile Navigation and Route Guidance

Automobile navigation systems develop and present to the driver various forms of navigation information useful in determining how to reach the desired destination.[5] Although navigation information may be perceived largely as a convenience by the individual driver, there are also broader societal needs according to a Federal Highway Administration study.[6] The study estimated that almost 7% of all distance traveled by noncommercial vehicles and over 12% of time spend driving is wasted because of poor navigation and route-following skills, thus contributing to congestion and accidents. The additional costs associated with this excess travel amount to $45 billion per year.

Most automobile navigation systems already on the market include an electronic road map display with icons indicating current vehicle location and the destination. More advanced systems also compute optimum routes and use simplified graphics and/or synthesized voice to provide real-time, step-by-step route guidance instructions for reaching the destination. Future systems may include "head-up" displays that project guidance or other information onto the windshield of the vehicle.

In addition to means for automatically and continuously determining vehicle location with sufficient accuracy to identify the road being traveled and each intersection approached, automobile navigation systems include digital road map databases giving the location, classification, and address ranges for each road in the areas where the vehicle operates. The map database may also include traffic regulations and typical travel times for individual road links for use in calculating optimum routes, as well as additional information such as the location and description of service stations, garages, parking, public buildings, hotels, restaurants, tourist attractions, and other types of directory information commonly used by travelers. Automobile navigation systems must also include man/machine interfaces for driver input and control and for presentation of navigation or route guidance information.

Figure 2 shows the major elements of a typical vehicular navigation system. Distance and heading (or heading change) sensors are invariably included for dead reckoning calculations, which, in combination with map matching, form the basic platform for keeping track of vehicle location. However, dead reckoning with map matching has the drawback of occasionally failing because of dead reckoning anomalies, extensive travel off mapped roads, ferry crossings, etc. Thus, the location sensor indicated by dashed lines is an optional means of providing absolute location to avoid occasional manual reinitialization when dead reckoning with map matching fails. Although electronic signposts or proximity beacons served to update vehicle location in some early developmental systems, most state-of-the-art systems include a GPS receiver for this purpose.

The data transceiver (see Fig. 2) is an option to permit navigation and route guidance systems to receive real-time traffic data from a traffic management center for use in determining optimum routes under prevailing conditions. The transceiver may also be used for transmitting measured link travel times to a traffic management center to maintain a real-time database of link travel times for transmittal to equipped vehicles. In-vehicle navigation and route guidance

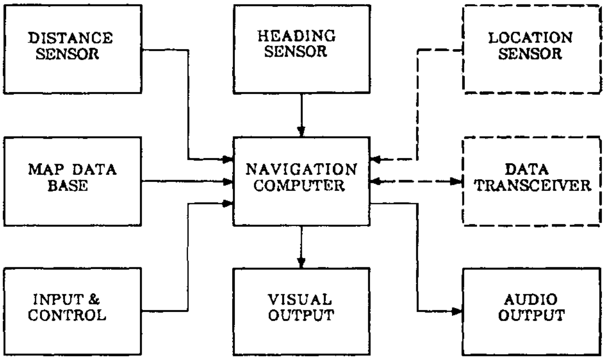


Fig. 2 Typical components and subsystems of vehicle navigation system. "Location sensor," if included, may be a GPS receiver or a proximity beacon receiver.[33]

systems coupled to traffic management centers by communication links are one of the most central aspects of IVHS.

C. Intelligent Vehicle Highway Systems

Intelligent vehicle highway systems apply computer, positioning, communications, and control technologies to integrate vehicles and highways in a coherent information network that facilitates the travel of individual vehicles, while optimizing traffic flow and increasing traffic capacity throughout the entire road system. Some basic concepts and components of IVHS are illustrated by Fig. 3.

Major efforts got underway in the United States, Europe, and Japan during the late 1980s to develop and apply IVHS to reduce congestion, improve mobility and road transportation efficiency, enhance safety, conserve energy, and protect the environment. IVHS AMERICA was formed in 1990 as a public–private educational and scientific organization to coordinate and promote the development of IVHS in the United States.

Although the taxonomy of IVHS is not yet fully consistent worldwide, the following six categories used for defining the United States' IVHS program encompass virtually all elements of IVHS approaches being pursued elsewhere as well:

1. Advanced Traffic Management Systems (ATMS)

Advanced traffic management systems extend real-time computer optimization of traffic signal timing to the urban road network level as opposed to individual

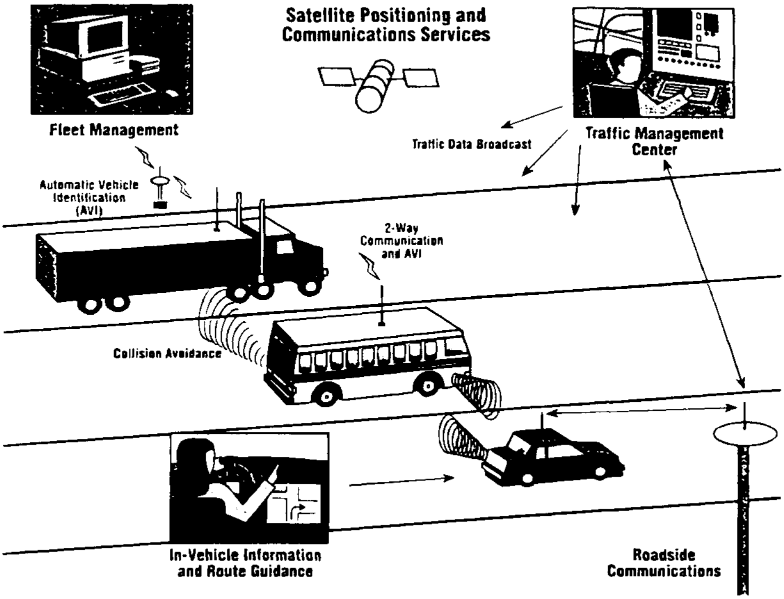

Satellite Positioning and Communications Services
Fleet Management
Traffic Management Center
Traffic Data Broadcast
Automatic Vehicle Identification (AVI)
2-Way Communication and AVI
Collision Avoidance
In-Vehicle Information and Route Guidance
Roadside Communications

Fig. 3 Elements of IVHS.[34]

intersections or streets. This requires information on traffic conditions throughout the network in a real-time database that may also serve as an information source for dynamic route guidance in advanced traveler information systems-equipped vehicles.

2. Advanced Traveler Information Systems (ATIS)

Advanced traveler information systems keep drivers informed of their location and provide route guidance to selected destinations along with information on services such as lodging, food, fuel, repair, medical facilities, etc. ATIS permit communication between in-vehicle equipment and ATMS for data on traffic conditions, diversion routes, alternative modes of transportation, etc. Although ATIS concepts originally centered on vehicular navigation and route guidance for drivers, new ATIS concepts include portable versions for use by pedestrians and multimodal travelers also.

3. Commercial Vehicle Operations (CVO)

These include vehicle tracking and fleet management systems for commercial and emergency vehicles to improve operational efficiency and increase safety. They also include technologies such as automatic vehicle classification (AVC), weigh-in-motion (WIM), and communications among automated regulatory checkpoints so that intercity trucks may travel among different jurisdictions with minimal stopping.

4. Advanced Vehicle Control Systems (AVCS)

These apply additional technologies to vehicles to detect obstacles and adjacent vehicles, thus enhancing vehicle control by augmenting driver performance. Advanced vehicle control systems assist in the prevention of collisions for safer high-speed driving to increase roadway capacity, and they will eventually interact with fully developed ATMS to enable automatic vehicle operation.

5. Advanced Public Transportation Systems (APTS)

In addition to applying the above IVHS technologies to public transportation systems, APTS have a strong foçus on customer interface. Examples include onboard displays (e.g., for next stop, transfer information, etc.), real-time displays at bus stops, and smart card fare systems as well as ride share and HOV (high-occupancy vehicle) information systems.

6. Advanced Rural Transportation Systems (ARTS)

These focus on issues and problems involving the development and application of IVHS to rural transportation. The major thrusts of ARTS include emergency communications (e.g., automated Mayday calls) and safety applications of IVHS technologies.

Intelligent vehicle highway systems are still in the early stages of development and, although numerous operational field trails are underway, relatively few actual applications of IVHS technology have been implemented. However, according to a General Accounting Office report,[7] available results indicate that travel times have been reduced by as much as 15%, and freeway speeds have been increased from 15 to 40 mph in congested areas. Other reported benefits include an 8–12% decrease in motor fuel consumption, an 8% decrease in hydrocarbon emissions, and an 18% decrease in roadway accidents.

Autonomous navigation systems started coming on the market in the late 1980s, and approximately 500,000 were already in use in Japan in 1993. ATIS versions with communication links to real-time traffic databases will start becoming available in the mid-1990s. An estimated 2.5 million vehicles per year will be sold with factory-installed ATIS by the year 2000, and 11 million per year will be sold by the year 2010, according IVHS AMERICA strategic planning projections.[8]

The IVHS movement also adds momentum to existing and emerging vehicle-tracking applications. Strategic planning for IVHS in the United States assumes that almost 300,000 CVO systems per year will be purchased by the year 2000. However, IVHS also has potential for modifying the architecture of vehicle-tracking systems when ATIS start becoming commonplace in all types of vehicles by the late 1990s. Because ATIS include navigation systems requiring accurate location information, it is likely that a common positioning subsystem will come to be used for both navigation and tracking.

II. Historical Background

Although vehicular navigation and tracking system concepts have become widely known only during the last few decades, their historical roots go much deeper.

A. Early Mechanical Systems

The world's first vehicular navigation system was the "south-pointing chariot," an automatic direction-keeping system developed by the Chinese around 200–300 AD (possibly earlier according to some legendary accounts), almost 1000 years before the magnetic compass was invented.[9] Its operation was based on the phenomenon that as a vehicle changes heading, the outer wheels travel farther than the inner wheels by a distance that is a simple mathematical function of the change in heading. When changing heading, a gear driven by the outer wheel of the south-pointing chariot automatically engaged and rotated a horizontal turntable to exactly offset the change in heading. Thus, a figure mounted on the turntable continuously pointed an outstretched arm in the same direction, like a compass needle, regardless of which way the chariot turned.

The differential odometer, the principle used in the south-pointing chariot, is now popular as a dead reckoning subsystem for modern automobile navigation systems because it is not subject to magnetic field disturbances as a magnetic compass is. It has the further advantage of not requiring additional hardware in automobiles equipped with wheel rotation sensors for antilock braking systems (ABS).

Mechanical route guides for automobiles began appearing in the United States around 1910 and were developed to aid drivers of early automobiles.[10] These pioneering devices incorporated the information of route maps in various forms including sequential instructions printed on a turntable, punched in a rotating disk, and printed on a moving tape, all being driven by an odometer shaft in synchronization with distance traveled along the route.

In fact, the odometer is itself a navigation device because it may be used in conjunction with road signs or road maps to estimate present position and monitor progress toward a destination. The mechanical route guides, however, went a step further to provide explicit real-time route instructions automatically at decision points along the way.

One of the earliest, and one of the most popular mechanical route guides to reach the market, was the Jones Live-Map, which was patented in 1909.[11] It consisted of a turntable slowly rotated by a gear train connected by flexible shaft to one of the vehicle wheels. Paper disks for individual routes with a scale of miles printed around their perimeters were mounted on the turntable beneath a glass cover with a fixed pointer. Detailed road directions keyed to specific distances from the beginning of a route came into view under the pointer at the time for execution. An advertisement for the Jones Live-Map claimed: "You take all the puzzling corners and forks with never a pause. You never stop to inquire. . . ."

The far more sophisticated Chadwick Road Guide was introduced in 1910.[12] Like the Live-Map, the Chadwick device rotated a calibrated disk in synchronization with distance traveled. However, the metal disk contained holes spaced to coincide with decision points along the route represented by the disk. An array of spring-loaded pins behind the slowly rotating disk was normally depressed, but when a punched hole traversed a pin, the pin released and raised a signal arm bearing a color-coded symbol indicating the action to be taken. Simultaneously, a bell sounded to draw the driver's attention to the signal. Moreover, beneath a pointer, the same disk gave printed information regarding the location or the action to be taken.

B. Early Electronic Systems

As the expanding roadway system became better marked with standardized signs, and reliable road maps became available, the need for route guidance devices diminished, and only sporadic developments occurred between World War I and the late 1960s. Mechanical approaches faded into the background with the advent of electrical servomechanisms, electronic controls, and, eventually, the digital computer.

One of the earliest land vehicle navigation systems to incorporate electronic elements was the vehicular odograph developed for military vehicles during World War II.[13] The vehicular odograph was a dead reckoning system that included a light beam and photocell arrangement to read the output of a magnetic compass. The compass output drove a servomechanism to rotate a mechanical shaft corresponding to vehicle heading. The shaft was coupled to a mechanical computer that resolved distanced traveled from an odometer shaft into X and Y components, and drove a stylus to trace the vehicle's course automatically on a map of corresponding scale.

In one of the first steps toward applications now called IVHS, the U.S. Bureau of Public Roads (now the Federal Highway Administration), started researching means in the late 1960s for integrating in-vehicle route guidance with traffic management under a project called ERGS (Electronic Route Guidance System).[14] This system used proximity beacons in the form of short-range transmitters with inductive loop antennas buried beneath the roadway at strategic intersections. A dash-mounted console with thumbwheel switches permitted the driver to enter a selected destination code that was transmitted when the vehicle unit was triggered by a roadside unit upon nearing key intersections. The roadside unit immediately selected the optimum route to the destination considering the current traffic patterns and transmitted instructions for display on the vehicle's console.

Although technically sound, ERGS required expensive roadside infrastructure, and the development was canceled in 1970, following limited testing at the subsystem level. Similar approaches were carried through further stages of development and testing during the 1970s in Japan and Europe. Although a quarter of a century old, the basic ERGS concept is still representative of IVHS systems that use proximity-beacons for communications and/or position updates. In the meantime, new communications approaches, availability of inexpensive onboard computers, development of map-matching algorithms, and the promise of satellite positioning gave rise to a number of alternatives to the proximity-beacon approach.

Another early development toward IVHS was map-matching of the early 1970s to augment dead reckoning in vehicular navigation systems. Networks of roads and streets may be modeled as internodal vectors in a digital map database, and a particular route may be programmed as a uniquȇ sequence of mathematical vectors. An onboard computer may be used to analyze dead reckoning inputs and match the deduced vehicle path with programmed routes to remove position discrepancies automatically that otherwise accumulate. Similarly, map-matching may be used to trace the location of a vehicle traveling arbitrary paths within a network.

The first map-matching system, called the automatic route control system (ARCS), was developed in 1971 for a commercial delivery operation rather than IVHS.[15] It used the differential odometer principle to compute the vehicle's approximate path. A map-matching algorithm correlated each sequentially measured (i.e., dead reckoned) route vector and change in heading with its map database counterpart. Thus, the vehicle's location was confirmed by ARCS, and real-time route guidance (albeit without consideration of real-time traffic conditions) was issued at appropriate points using prerecorded voice messages. A second version issued route instructions visually, using simplified graphics on a plasma display panel.

III. Enabling/Supporting Technologies

Positioning technologies are fundamental requirements of both vehicle navigation systems and vehicle-tracking systems. Almost all vehicular navigation systems are heavily dependent upon dead reckoning with map matching as the main positioning technology, but most such systems now include GPS to good advantage. Map matching as well as route guidance requires digital road maps.

Another important supporting technology is mobile data communications. Although vehicle-tracking systems may not require digital road maps aboard the vehicle, they do require a mobile data communications link between the vehicle and dispatch office unless they use infrastructure-based integrated positioning and communications technologies. However, GPS and other radiopositioning technologies, especially Loran-C, are typically the sole or main technologies used for vehicle-tracking systems.

A. Dead Reckoning

Dead reckoning, the process of calculating location by integrating measured increments of distance and direction of travel relative to a known location, is used in many vehicle-tracking systems and in virtually all vehicle navigation systems.

1. Mathematical Formulation

The basic mathematical formulation for determining a vehicle's current coordinates, X_n and Y_n, relative to its initial coordinates, X_0 and Y_0, as depicted in Fig. 4 is as follows:

$$X_n = X_0 + \sum_{1}^{n} \Delta X_i = X_0 + \sum_{1}^{n} \Delta \ell_i \sin \phi_i$$

$$Y_n = Y_0 + \sum_{1}^{n} \Delta Y_i = Y_0 + \sum_{1}^{n} \Delta \ell_i \cos \phi_i$$

where ϕ_i is the heading associated with ℓ_i, the ith increment of travel, as illustrated. Thus, dead reckoning for vehicles requires a means for sensing distance traveled and heading (or change in heading).

2. Distance/Speed Sensors

Distance measurements for vehicle navigation systems are usually made with an electronic version of the odometer. Electronic odometers provide discrete

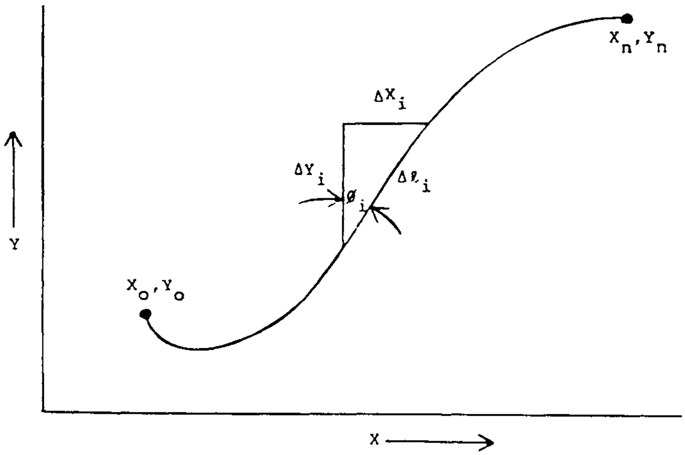


Fig. 4 Dead-reckoning formulation.

signals from a rotating shaft or wheel, and a conversion factor is applied to obtain the incremental distance associated with each signal. Magnetic, inductive, or capacitive sensors typically are mounted on a stationary member to detect the passage of closely spaced magnets or metallic protrusions attached to the hub or rim of a vehicle's wheel. Automobiles equipped with ABS already have wheel sensors that may also be used with navigation systems.

Odometer measurements are subject to a number of random and systematic errors, some of which can be defined and corrected in the distance conversion process. The difference in the diameter of a new tire and a well-worn tire, for example, can contribute distance errors as high as 3%. The error in distance measurements increases by approximately 0.1–0.7% when vehicle speed is increased by 25 mph because of the effect of centrifugal force on the tires, and a 10 psi change in tire pressure can induce an error of 0.25–1.1% percent.[16]

3. *Heading/Heading-Change Sensors*

Vehicle heading may be determined by direct measurement with a magnetic compass, or by keeping track of heading relative to an initial heading by accumulating individual changes in heading. A number of alternative means are available for measuring vehicle heading or heading changes. However, most have at least one drawback. As a result, it is a common practice for two different types of sensors to be used in combination to offset one another's weaknesses.

The magnetic compass's well-known accuracy problems caused by anomalies in the Earth's magnetic field are compounded when installed in a vehicle because of induced fields that depend upon vehicle heading. In addition, a vehicle may have a strong permanent magnetic field of its own, and subpermanent magnetism may be acquired or lost when hitting bumps. Buildings, bridges, and other structures external to a vehicle can also cause magnetic aberrations. Moreover, unless a compass is pendulum mounted, more errors are introduced when operating on an incline. A special consideration in installing magnetic compasses in trucks is the potential for disturbances from various types of cargo.

Nevertheless, some form of magnetic compass is used in most vehicular navigation systems. Compact, solid-state, flux-gate compass with software processes for compensating errors resulting from both permanent and induced magnetism of the vehicle[17] are especially popular in current systems. In many applications, both a flux-gate compass and a differential odometer or a gyroscopic device are used along with a software filtering process that combines the two outputs. Relatively more weight is placed on the differential odometer or gyro output for short-term changes in heading and on the compass for longer-term trends in absolute heading.

A differential odometer essentially is a pair of odometers, one each for the wheels on opposite ends of an axle. When a vehicle changes heading by an amount θ, as illustrated in Fig. 5, the outer wheel travels farther than the inner wheel by ΔD:

$$\Delta D = D_R - D_L$$

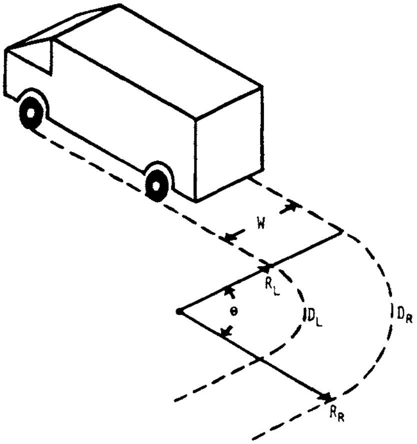


Fig. 5 Differential odometer principle.

Expressed in terms of heading change and vehicle width *W*, the difference in wheel travel is:

$$\Delta D = \theta(R_R - R_L) = \theta W$$

Thus, $\theta = \Delta D/W$.

Most other types of dead reckoning sensors used to detect heading changes for vehicle location and navigation systems use some form of the gyroscopic principle. These range from traditional spinning devices and gas-jet sensors to vibrating bars and fiberoptic gyros, and, although more expensive than differential odometer sensors, they are much simpler to install. The fiberoptic gyro started appearing in production automobile navigation systems in 1991.[18]

B. Digital Road Maps

An onboard digital road map database is an essential feature of vehicular navigation and route guidance systems. Digital maps also play important roles in vehicle-tracking systems for displaying vehicle location at the fleet dispatch office. The two basic approaches to digitizing maps are matrix encoding and vector encoding. Matrix encoding preserves map detail at a selected degree of resolution in the form of a digitized image, whereas vector encoding models roads in a mathematically abstracted form.

A matrix encoded map may be thought of as a digitized image in which each image element or pixel, as determined by an *X*–*Y* grid with arbitrary spacing, is defined by digital data giving characteristics such as shade or color. Thus, in addition to requiring more data storage, matrix encoding does not facilitate analytical treatment (such as map matching or route finding) of logical connections among the road elements. Nonetheless, matrix encoding is sometimes used in digitizing maps for display purposes such as tracking systems that superimpose vehicle location or other information.

The vector encoding approach applies mathematical modeling concepts to represent such geometrical features as roadways and boundaries with a minimum of data. By considering each road or street as a series of straight lines and each intersection as a node, a map may be viewed as a set of interrelated nodes, lines, and enclosed areas, as illustrated in Fig. 6.[19] Nodes may be identified by their coordinates (latitude and longitude). Additional nodes, typically called "shape points," are positioned along curves if the link between two intersections is not a straight line. Thus, curves are approximated by a series of vectors connecting shape points, whereas a single vector directly connects the node points representing successive intersections if there are no curves on the road segment in between.

The *X–Y* coordinates of node points may be encoded from maps or aerial photographs. In practice, they are usually encoded using special work stations that record the coordinates of a given point when the crosshair of an instrument is placed over the point and a button pressed. This process has been automated in varying degrees. In some cases, the printed map is scanned to obtain a matrix image, which is then converted to vector form by software.

Various combinations of attributes associated with the encoded road network are included in digital map databases. Of particular importance are roadway classifications, street names, and address ranges between nodes. Map databases used with systems that give turn-by-turn route guidance instructions also require traffic attributes such as turn restrictions and delineation of one-way streets. Directory and "yellow pages" information for selecting attractions, parking, restaurants, hotels, emergency facilities, etc. commonly are included.

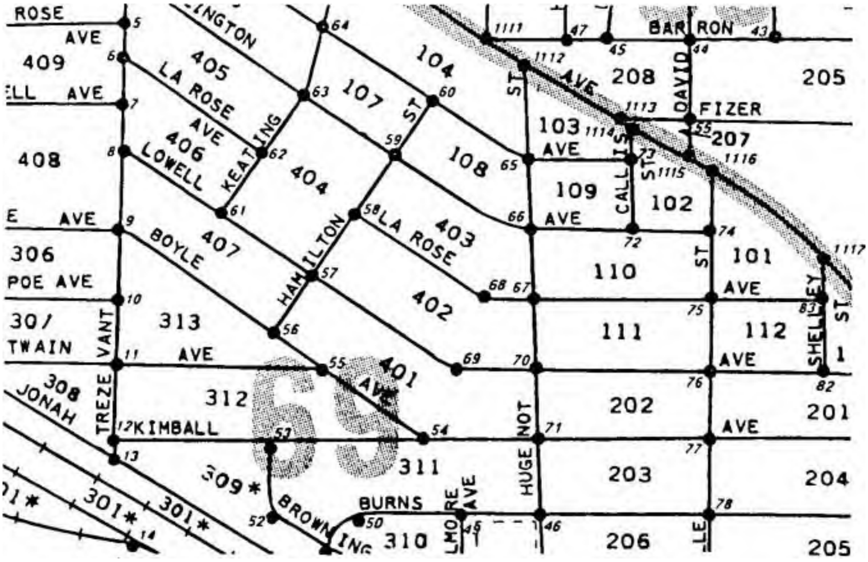


Fig. 6 Nodes and street segments of vector encoded map.[19]

C. Map Matching

The fact that motor vehicles are largely constrained to a finite network of streets and roads with only occasional excursions into driveways, parking lots, etc. makes it possible for computer algorithms utilizing road map information stored in vector encoded form to correlate a vehicle's path approximated by dead reckoning or other means with the digital map to maintain accurate knowledge of the vehicle's location within the defined road network. Known as "map matching" this process is used in virtually all vehicle navigation and route guidance systems. Most map matching algorithms may be classified as either semideterministic or probabilistic.[20]

1. Semideterministic

Semideterministic map-matching algorithms assume that the equipped vehicle is essentially confined to a defined route or road network, and thus are designed to determine where the vehicle is along a route or within the network. The basic concept can be illustrated by tracking the location of a vehicle over the simple route shown in Fig. 7a.

Figure 7b defines the route from node A through nodes B, C, D, E, and thence back to A in terms of instantaneous direction, ϕ, of travel vs cumulative distance, L, from the beginning. Locations of nodes where direction changes occur (or could occur) thus are defined in terms of distance L. The solid line is the plot of heading vs distance corresponding to the simple route. Alternative routes emanating from each node are indicated by dashed lines.

The kernel of a semideterministic map-matching algorithm is shown in highly simplified form in Fig. 8. Once initialized at a starting location ($\phi = 90$ deg and $L = 0$ at Node A in the simple example of Fig. 7), the algorithm, in effect, repeatedly asks, "Is the vehicle still following the route?" and "What is the present location along the route?" The vehicle is confirmed on the route if certain tests are satisfied. The location along the route is estimated by odometry, and error in the estimate is automatically removed at each node where it is determined that an expected change in the vehicle heading actually occurs.

The simplified deterministic map-matching algorithm is driven by interrupts from differential odometer sensors installed on the left and right wheels. The distance L from the beginning of a route segment is updated by adding an increment ΔL for each left wheel interrupt, and the vehicle heading ϕ is updated by adding an increment $\Delta\phi$ calculated from the difference in travel by the left and right wheels occurring since the count N was last set to 0. As explained below, the N counter controls monitoring for unexpected heading changes occurring over relatively short distances.

Unless the turn flag is set to denote that the vehicle is approaching a distance L where a heading change should occur, count N is checked after each interrupt to determine if it has reached a limit C corresponding to an arbitrary amount of travel on the order of several meters. When the count limit C is reached, a test is made to determine if ϕ is within arbitrary limits (say ± 5 or 10 deg to allow for lane changes, slight road curvature, etc.). If so, ϕ is reset to ϕ_0 (the initial direction of the vector being traveled) and N is set to 0 to start another cycle of monitoring for unexpected heading changes.

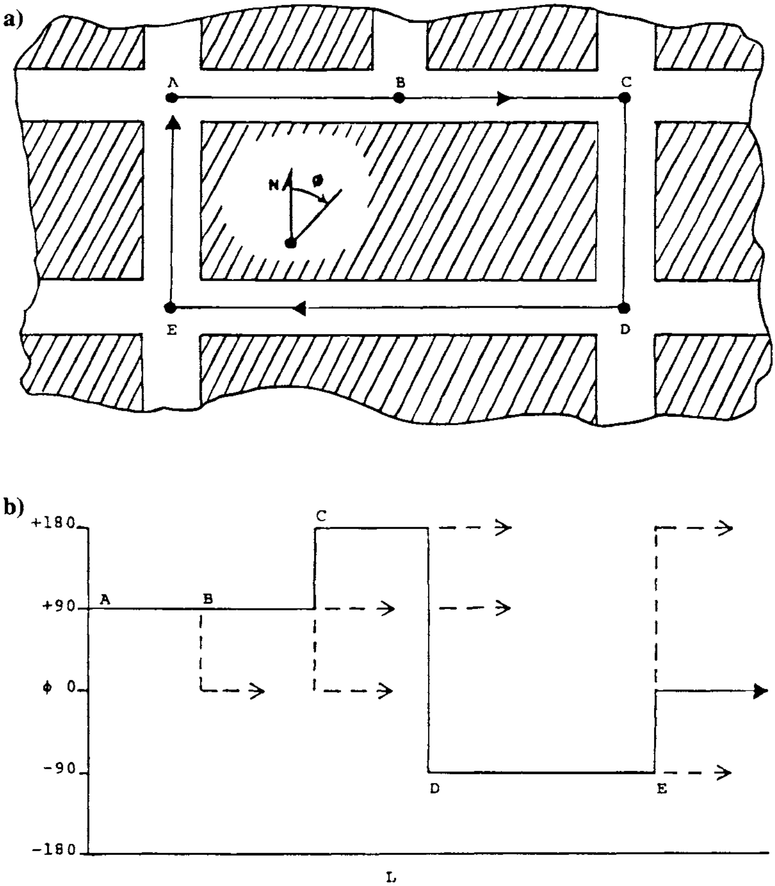


Fig. 7 Simplified route and vector model.[20]

Thus, in addition to verifying that the vehicle remains on the route while traveling between nodes, this process removes error in measured vehicle heading that accumulates while $0<N<C$. If the above test finds ϕ to be outside the limits, the vehicle is presumed to have turned off the route (perhaps into an unmapped driveway or parking lot), and other routines are called into play. For example, the driver could be informed of a route error and issued recovery instructions.

When the vehicle approaches within an arbitrary distance (say 75 m) of a node where a change in vehicle heading should occur, the turn flag is set and a route guidance instruction indicating the direction of the turn and, if appropriate, the name of the road to take is issued. The algorithm then continuously monitors changes in ϕ to confirm that the midpoint of the expected turn is reached within arbitrary limits (say 10 m) of the value of L specified for the node, and to confirm that the turn is completed.

When the midpoint is reached, the current value of L is adjusted to that specified, thus removing any error in the measured distance accumulated since

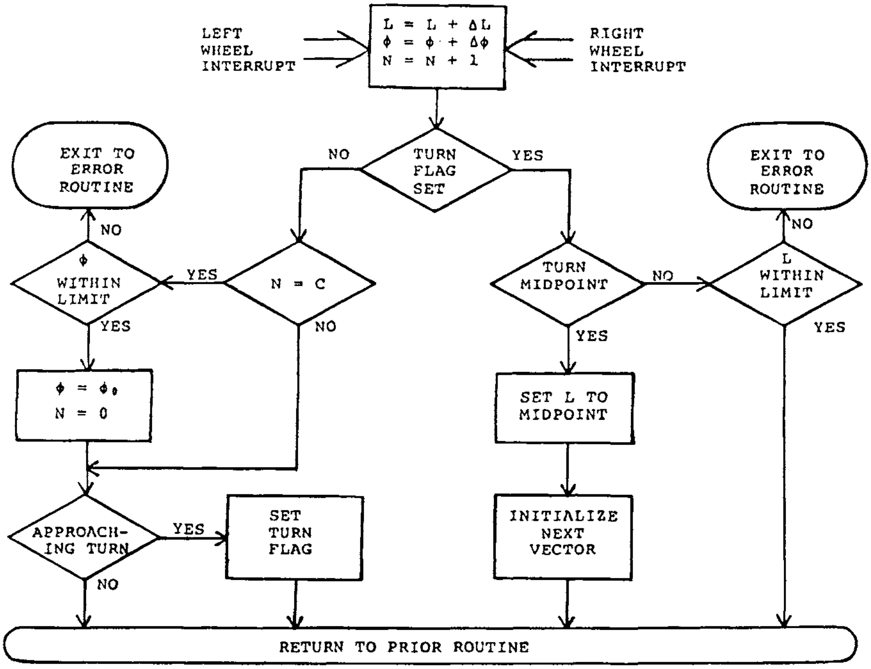


Fig. 8 Simplified map-matching algorithm.[20]

the last turn. If the expected turn is not observed within the allowed limits on distance *L*, the vehicle is assumed to have missed the turn or to have taken an alternate turn (see dashed lines in Fig. 7b), and other routines may be entered to identify the alternate route taken from the node.

The semideterministic algorithm concept outlined in the preceding paragraphs may be extended to tracking a vehicle's location as it moves over arbitrary routes within a road network rather than following a preplanned route. As long as the vehicle stays on roadways defined by a vector-encoded digital map, the vehicle must exit each node via some vector. Thus, a map-matching algorithm can identify successive vectors traveled by measuring the direction of vehicle travel as it leaves each node and comparing the vehicle direction with that of various vectors emanating from the node.

2. *Probabilistic*

Another type of map-matching routine is required for tracking vehicles not presumed to be constrained to the roads of a particular route or network. When the vehicle departs from the defined route or road network, (e.g., into a parking lot), or appears to depart as a result of dead reckoning error, the routine repeatedly compares the vehicle's dead reckoned coordinates with those of the links surrounding the off-road area that encompasses the vehicle location in order to recognize where the vehicle returns to the road network. Unlike travel on defined roadways, frequent map-matching adjustments do not prevent the accumulation of dead reckoning error. Thus, depending upon the distance traveled off road and the accuracy of the dead reckoning sensors, there may be considerable uncertainty in vehicle coordinates, which could produce misleading conclusions when tested against the surrounding links.

Probabilistic map-matching algorithms are used to minimize the potential of off-road errors by maintaining a running estimate of uncertainty in dead reckoned location, which is taken into consideration in determining whether the vehicle is on a street or not. The estimate of location uncertainty is reduced each time it is deemed that the vehicle is on a street, but the uncertainty resumes growth in proportion to further vehicle travel until the next match occurs. Thus, a probabilistic algorithm repeatedly asks, "Where is the vehicle?" with no a priori presumption that it is on a road.

D. Integration with GPS

Just as backing up GPS with dead reckoning may be highly useful in some vehicle-tracking applications, GPS may be integrated with map matching in others. Backing up dead reckoning and map matching in vehicular navigation systems with GPS eliminates the occasional failure of the former. Although failures may be infrequent (e.g., *Travelpilot,* Sec. IV.A), manual reinitialization can be a tedious and time-consuming task. Thus, GPS augmentation is commonly used in state-of-the-art automobile navigation systems.

The integration of GPS with dead reckoning as is often done for vehicle-tracking systems may be accomplished by a variety of approaches ranging from very simple to relatively complex. In the simplest case, dead reckoning may be used to fill in discontinuities in GPS position by initializing the dead reckoning position to the last GPS fix and incrementing the position as outlined in Sec. III.A. At the other extreme are sophisticated software filter approaches, such as the example outlined in Sec. IV.D.

GPS receivers may be integrated with dead reckoning and map matching in automobile navigation systems through relatively simple modifications of map-matching algorithms or through use of Kalman or other software filtering schemes. In the simplest approach, the GPS position may be ignored if it is within reasonable agreement with the dead-reckoning/map-matching position. However, when combining positions from the different techniques to determine the most likely location, the map-matching software must take into account the probable errors or uncertainties of each of the different techniques. This can be accomplished by incorporating filters in map-matching software.

The Kalman filter is the best known filter technology used for combining by developers of vehicle-tracking and navigation software. The basic mathematical models from which a Kalman filter equations are derived are fk at time epoch tk for map matching, $fk + 1$ at time epoch $tk + 1$ for GPS, and $gk, k + 1$ for the interval between map-matching and GPS updates.[21] A Kalman filter uses the predicted state vectors and covariance matrices to compute the filtered state vectors and covariance matrices. This process is repeated for each time epoch.

E. Mobile Data Communications

Vehicle-tracking systems as well as navigation systems integrated with IVHS depend upon mobile communication for exchanging information between in-vehicle equipment and dispatch offices or traffic control centers. The principal communication systems identified for electronic transmission of information to and from vehicles are listed and broadly characterized in Table 1.[22]

Regulatory bodies allow commercial FM and TV broadcast stations to transmit inaudible ancillary information on sideband channels displaced 53–99 KHz from the central frequency. Called subcarrier authorization (SCA) in the United States, these sideband channels may carry analog or digital signals that are detected and decoded by special attachments or design features of ordinary receivers. The European Radio Data System (RDS) standardized by the European Broadcasting Union in 1984 operates on the same principle for transmitting station and programming data for automatic tuning along with other information in digital form.[23]

Table 1 Characteristics of alternative mobile communication approaches

Approach	Characteristics
Broadcast SCA	One-way only Voice and data Low data rates Extended area coverage Includes RDS
Proximity beacon	One-way or two-way Data only High data rates Spot-area coverage
Inductive loop	One-way or two-way Data only Low data rates Spot-Area coverage
Land-mobile radio (dedicated)	Two-way Voice and data Local-area coverage
Specialized mobile radio (SMR)	Two-way Voice and data Extended-area coverage
Cellular radio	Two-way Voice and data Local/extended-area coverage
Mobile satellite communications	One-way or two-way Voice or data Wide-area coverage
Meteor burst communications	Two-way Data only Wide-area coverage Involves time delays

Source: Ref. 22.

Use of the SCA approach for area broadcast of digitized information to vehicles requires relatively little additional equipment other than a minor enhancement of the FM radio receiver commonplace in vehicles. However, broadcast SCA has the disadvantage of being limited to one-way communication into vehicles, thus it cannot be used for vehicle tracking.

Infrared and microwave proximity beacons, which have the advantages of high data rates in combination with spot coverage (i.e., typically limited to tens of meters) that permit messages to be tailored to highly localized needs, are strong candidates for a mobile data communications role in route guidance systems being developed and tested in Europe and Japan. In these densely populated geographic areas, heavy infrastructure costs of installing beacons at close intervals throughout the road system and integrating them into overall communications networks do not seem as formidable as in the United States where, except for use as electronic signposts in vehicle-tracking systems for transit buses, proximity beacons have been used for few applications.

Inductive loops, which essentially are proximity beacons in the form of radio antennas buried beneath the roadway, have the low data rates characteristic of radio-frequency proximity beacons. They have the further disadvantage of being very expensive and awkward to install because traffic must be temporarily diverted from lanes being equipped. Maintenance is also expensive because of the wear and tear from traffic. Nonetheless, inductive loops were considered for communications in the early ERGS route guidance system research (see Sec. II.B), and they are being re-examined under current IVHS programs.

Although individually owned and operated land mobile radio systems for two-way voice communication have seen widespread use for fleet management within city areas, their spread has been severely limited by the amount of spectrum allocated for this purpose. However, great gains in dedicated land mobile throughput have resulted from increased use of data communications as an alternative to voice.

Specialized mobile radio is a class of land mobile radio service first authorized by the FCC in 1974. Specialized mobile radio quickly emerged to become the pre-eminent provider of private land mobile communications service, particularly in large metropolitan areas. Basically, SMR is a business and regulatory approach that permits different entities to share common transmission facilities and frequencies. With one type of trunked SMR service, a user vehicle unit monitors a number of frequency channels for a unique digital code identifying a transmission addressed to the individual unit and monitors channel use to select an available frequency for transmitting its own messages addressed to the user's dispatch station. With the second type of trunked system, the vehicle unit monitors a "control channel" that manages channel assignments.

Cellular radio telephones have provided a major step forward for voice communications in the vehicular environment, and are now seeing increased use for data communications. However, the present scheme of using cellular radio technology exclusively for full-duplex individually addressed communications will limit the potential use of cellular radio in providing traffic and routing data for vehicular navigation and information systems. Effective adaptation of cellular radio to the data communications requirements of navigation and route guidance would require the establishment of a dedicated channel for repeatedly broadcasting the map updates, traffic data, etc., by all cell transmitters in the local area served by a cellular system. Equipped vehicles could have special receive-only units to detect the data for transfer to the onboard equipment.

In the meantime, commercial fleet management applications of conventional cellular telephone are growing, and have particular attraction for small vehicle fleets with only infrequent requirements for voice communications within local cellular service areas. Cellular data links are also being considered by a number of vehicle location monitoring systems integrators for localized applications.

Unlike urban vehicle location monitoring, which can use short-range land-mobile radio or cellular telephone communication links, crosscountry truck location monitoring requires long-distance communication links such as those characteristic of mobile satellite services. Thus, satellite communications have long been viewed as having great potential for crosscountry trucking applications because of their wide-area coverage. A well-established example is the Omni-TRACS two-way satellite communication and position reporting service offered by Qualcomm, Inc. using geostationary satellites. Low-Earth-orbit (LEO) satellite systems such as Iridium proposed by Motorola, Inc. and ORBCOMM proposed by Orbital Communications Corp. also hold potential for extending mobile communications in rural and remote areas. However, the wide area covered by satellites is not an advantage for traffic data communications because traffic data are not useful outside the subject local area.

IV. Examples of Integrated Systems

The following examples of state-of-the-art systems collectively illustrate the recent evolution and present trends worldwide for vehicular navigation, route guidance, and tracking systems.

A. Etak Navigator™/Bosch Travelpilot™

The first commercially available automobile navigation system to include digitized road maps, dead reckoning with map matching, and an electronic map display was the Etak, Inc. Navigator introduced in California in the mid-1980s.[24] It used a flux-gate compass and differential odometer for dead reckoning. The equivalent of two printed city street maps were vector encoded and stored on 3.5-Mb digital cassettes for map matching and display purposes. Although sales were modest, the highly publicized Etak Navigator drew widespread attention to the concept of an electronic map display with icons showing current location and destination.

The Travelpilot essentially is a second generation of the Navigator jointly designed by Etak, Inc. and Bosch GmbH.[25] It was introduced in Germany in 1989 and in the United States two years later. One of its most conspicuous enhancements is the use of CD-ROM storage for digitized maps. The 640-Mb capacity permits the entire map of some countries to be stored on a single CD-ROM.

The primary function of the Travelpilot vehicle navigation and information system is to display to the driver a road map of the area around the vehicle, as illustrated in Fig. 9. The vehicle location and heading are indicated by an arrowhead icon below the center of the screen. The vertical bar at the right edge of the map indicates the display scale that can be zoomed in to 1/8 mile for complete street detail or out to 30 miles to show only major highways. The map is normally oriented so that the direction in which the vehicle is heading points straight up on the display, thus allowing the driver to relate the map display easily to the view outside. However, the map may be viewed in a north up orientation for reference purposes when parked.

Also when parked, a menu accessible through the "MEN" button permits use of soft-labeled buttons in a speller and scroller scheme to enter destinations by street address, intersection, etc. Travelpilot uses a process called "geocoding" to locate an input destination and display it as a flashing star on the map. As illustrated in Fig. 9, a destination geocoded by street address is bracketed by two flashing stars when the map is zoomed in. In this case, the stars mark the block whose address range includes the street number of the destination. A line of information across the top of the map display indicates the crow-flight distance and points the direction from the vehicle's current location to the destination. Up to 100 input destinations may be stored for future use.

A submenu provides several methods for the driver to reset the vehicle's position on the map if the Travelpilot gets off-track. The frequency with which the system requires reinitializing depends upon dead-reckoning anomalies and the completeness and accuracy of the map data for the area being driven. For example, map matching typically fails once in a thousand miles when operating in an environment such as greater Los Angeles or Dallas/Fort Worth. As for location accuracy the rest of the time (i.e., with map-matching operative), Travelpilot is claimed to have infinitesimal error relative to the map. The map-matching performance is compared to that of a servo-amplifier in which map-matching failure corresponds to loosing servolock to the map.

In addition to the CD-ROM player and vector-drawn 4.5-in. monochromatic display, the Travelpilot hardware includes a V50 processor, 1/2 Mb DRAM, 64-

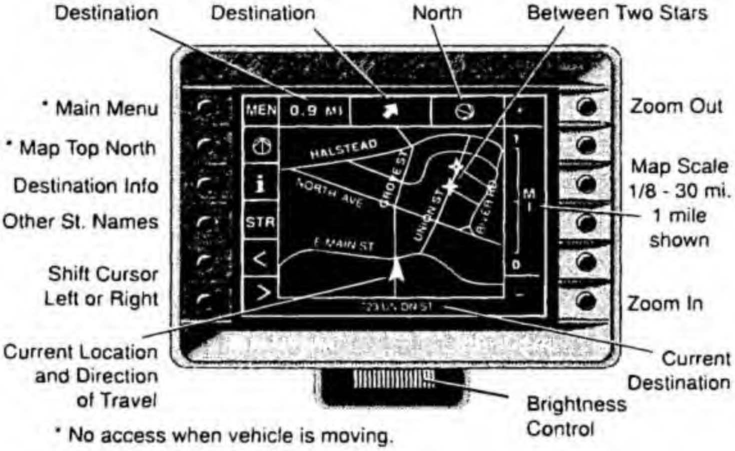


Fig. 9 Bosch Travelpilot display and controls.

Kb EPROM, and 8-Kb nonvolatile RAM. The nonvolatile RAM is used for storing vehicle location while the ignition is off, calibration factors, up to 100 saved destinations, etc.

The Travelpilot may interact with other devices through a RS-232 serial port and an expansion card slot. For example, Travelpilots in 400 Los Angeles fire trucks and ambulances are connected by digital packet radio to the city's emergency control center. The emergency operators can monitor each vehicle's location and status, and can send destinations directly to a vehicle's Travelpilot for emergency dispatch.

B. Toyota Electro-Multivision

Introduced in 1987 as the first sophisticated navigation system available as a factory option on automobiles sold in Japan, the Electro-Multivision has undergone numerous refinements including the recent addition of routing and voice guidance features. Except for a few features, it is representative of the more comprehensive models of navigation systems now available in Japan from almost all of the major automobile and electronics manufacturers.

The design and features of the original version of the Electro-Multivision can be summarized with reference to those of the Travelpilot. Both use dead reckoning and digitized maps stored on CD-ROM for display on a CRT screen with an icon representing present position, and are generally similar in their basic navigation features. However, a raster-scan color CRT rather than a vector-drawn monochromatic CRT is used in the Electro-Multivision. Also unlike Travelpilot, the Electro-Multivision map database includes "yellow pages" information such as the locations of facilities likely to be of interest to motorists.

The Electro-Multivision may also be used as a reference atlas. An initial display shows a color map of all Japan with 16 superimposed rectangles. Touching a particular rectangle causes the map area it encompasses to zoom and fill the entire screen, again with grid lines superimposed to form 16 rectangles. Thus, a few touches of the screen takes the driver from an overview of the entire country down to major roads and landmarks in some quarter of Tokyo. However, in spite of Electro-Multivision's sophisticated map-handling capabilities, map matching was not used in the first version because the digital maps then available for Japan did not contain sufficient detail at the city street level.

In addition to detailed digital maps and map matching, subsequent versions of Electro-Multivision include a GPS receiver and a color LCD rather than CRT display.[26] In 1991, a routing feature was added to calculate a suggested route to specified destinations and highlight the trace on the LCD map display.[27] The most recent version[28] adds synthesized voice route guidance instructions.

As with most other state-of-the-art automobile navigation systems in Japan, the Electro-Multivision navigation features are integrated with a full suite of entertainment features (e.g., AM-FM radio, tape cassette, audio CD player, color TV, etc.). In addition, Electro-Multivision includes a CCD camera for rear-vision on the LCD screen. Figure 10 shows the typical layout of Electro-Multivision components installed in an automobile.

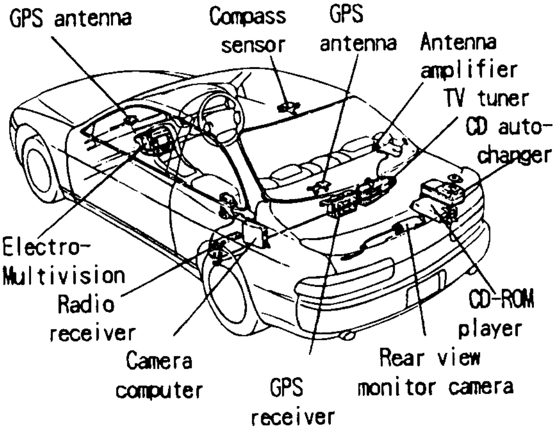


Fig. 10 Distribution of Toyota Electro-Multivision system elements in automobile.[27]

C. TravTek Driver Information System

Unlike Travelpilot and Electro-Multivision which are automobile systems already available in certain markets, TravTek is a functional prototype of a navigation-based in-vehicle traveler information system developed specifically for the TravTek IVHS operational field trial conducted for a one-year period ending March 31, 1993 in Orlando, Florida. The field trial was a joint public sector–private sector project with the primary objective of obtaining field data on the acceptance and use by drivers of navigation and other information provided by comprehensive in-vehicle systems linked with traffic operations and other data centers.

General Motors equipped a total of 100 cars with the system shown schematically in Fig. 11 to provide navigation, route selection and guidance, real-time traffic information, local "yellow pages" and tourist information, and cellular phone service.[29] Most of these vehicles were made available to Orlando visitors through Avis Rent A Car for short-term trials and the rest were assigned to local drivers for extended periods. The American Automobile Association selected the test subjects and operated a TravTek Information and Services Center that could be accessed via cellular telephone.

The City of Orlando, in conjunction with the Federal Highway Administration and the Florida Department of Transportation, operated a supporting Traffic Management Center that consolidated traffic data from various sources including "probe" data consisting of road segment travel times received from the equipped vehicles themselves. Data communications between the equipped vehicles and the Traffic Management Center were via specialized mobile radio (SMR).

TravTek navigation is based on a combination of dead reckoning and map matching with a GPS receiver playing a "watchdog" role. Although functionally realistic, the TravTek in-vehicle system design made extensive use of readily available modules that would not typically appear in a production system. For example, rather than consolidated databases stored on CD-ROM, TravTek used

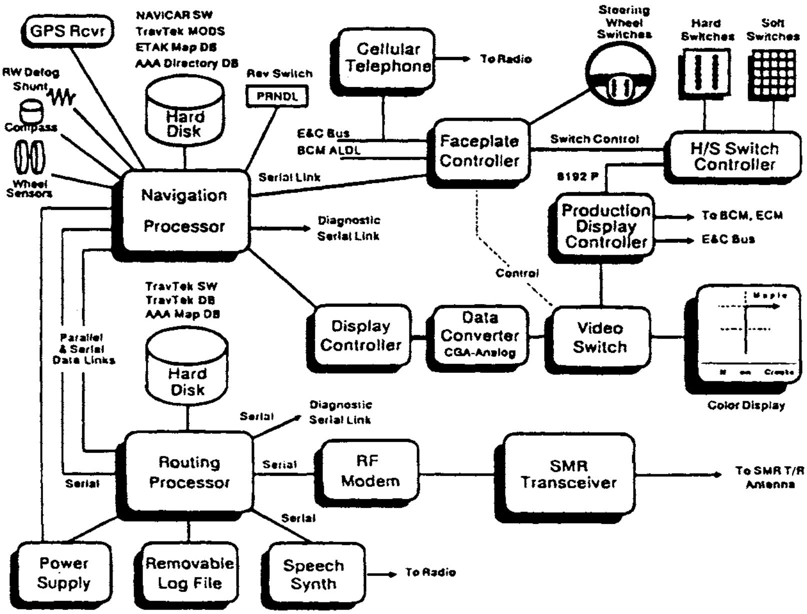


Fig. 11 Architecture of TravTek vehicle system.[29]

separate map databases stored on separate hard disk drives for navigation processing and route guidance processing.[30]

Similar to Travelpilot and Electro-Multivision, the TravTek navigation function superimposes vehicle location and destination on the map display screen and, like Electro-Multivision, highlights suggested routes on the map display and issues route guidance instructions via synthesized voice. In addition, turn-by-turn route guidance in the form of simplified graphics may be displayed, as indicated by Fig. 11.

Based on analysis of questionnaires completed by some 3000 test subjects, TravTek's effectiveness and benefits are rated very highly by drivers (5.1 on a scale of 1.0–6.0).

D. NavTrax(TM) Fleet Management System

NavTrax is a dispatch-type automatic vehicle location reporting system developed by Pulsearch Navigation Systems of Calgary, Alberta, Canada.[31] The positioning module is a robust GPS-based system integrated with dead-reckoning devices by a decentralized-federated filter making the module fault tolerant and suited for off-road as well as on-road use. The dispatch center subsystem provides map displays, means for selecting and polling (i.e., sending periodic requests for vehicles to report their positions), logging of fleet movements, etc. To date, the vehicle and the dispatch center have been linked by two-way UHF/VHF communications, although plans are underway to implement cellular technology as well.

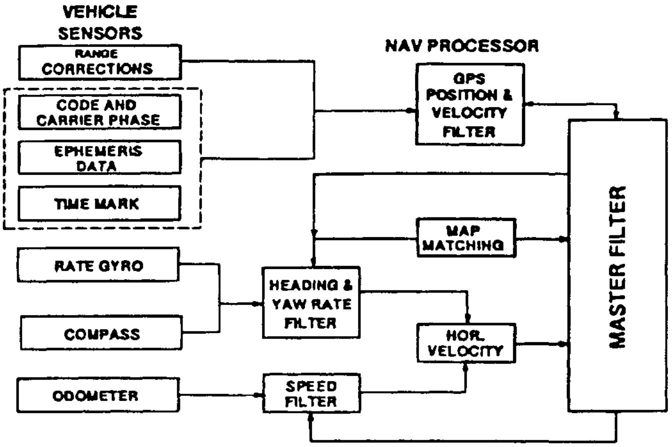


Fig. 12 Navigation module of NavTrax fleet management system.[32]

The major elements of the NavTrax positioning module are shown in Fig. 12.[32] Real-time position coordinates are continuously computed from sensor inputs including GPS pseudorange and carrier phase rate for position and velocity; rate gyro for azimuth change; compass for azimuth; and odometer for speed. With the decentralized-federated filtering approach, the reference filter is GPS based, while each dead reckoning sensor has its own local filter for determination of biases and fault detection. The reference filter and all local filters feed into the master filter where fusion of all position information takes place. Fusion feedback from the master filter to each local filter provides for frequent automatic calibration updates for each sensor.

The NavTrax filter scheme makes maximum use of short bursts of GPS information as it becomes available at intersections and openings in tree-covered areas. However, positioning can be continued for extended periods on the basis of dead-reckoning information alone. During periods of long lapses of GPS information, a limited form of map matching is carried out by taking the dead-reckoning coordinates and entering the road network database and exiting with location in the database frame of reference. The location is based on a probability function using the most probable road segment considering proximity, azimuth, and past connectivity. The associated map coordinates and azimuth of the road link are used in the filter, along with their associated variance information, to determine the best position of the vehicle.

NavTrax has been carried through several stages of Beta testing and is now being marketed. The trial applications included Calgary police vehicle dispatch, Amoco oil field operations, and vehicle location monitoring by the Royal Canadian Mounted Police.

References

[1] French, R. L., "Land Vehicle Navigation—A Worldwide Perspective," *Journal of Navigation*, Vol. 44, No. 1, 1991, pp. 25–29.

[2]"Automatic Vehicle Monitoring Program Digest," Urban Mass Transit Administration Report DOT-TSC-UMTA-81-11, April 1981.

[3]Rothblatt, M. A., *Radiodetermination Satellite Services and Standards,* Artech House, Boston, 1987.

[4]Sellers, D. L., and Bernard, T. J., "An Update on the OmniTRACS Two-Way Satellite Mobile Communications System and its Application to the Schneider National Truckload Fleet," *Proceedings of the 1992 International Congress on Transportation Electronics,* Society of Automotive Engineers, Dearborn, MI, 1992 (SAE P-260), pp. 351–356.

[5]French, R. L., "The Evolving Roles of Vehicular Navigation," *Navigation,* Vol. 34, No. 3, 1987, pp. 212–228.

[6]King, G. E., "Economic Assessment of Potential Solutions for Improving Motorist Route Following," Federal Highway Administration Rept. FHWA/RD-86/029, 1986.

[7]"Smart Highways: An Assessment of Their Potential to Improve Travel," U.S. General Accounting Office Rept. GAO/PEMD-91-18, May 1991.

[8]"Strategic Plan for Intelligent Vehicle-Highway Systems in the United States," IVHS America Rept. IVHS-AMER-92-3, May 1992.

[9]Needham, J., "Science and Civilization in China," *Mechanical Engineering,* Vol. 4, Part II, Cambridge University Press, New York, 1965.

[10]French, R. L., "U.S. Automobile Navigation: Early Mechanical Systems," *Navigation News,* Vol. 4, No. 3, 1989, pp. 6–7.

[11]Perry, H. W., "Some Remarkable Mechanical Road Guides," *Scientific American,* Vol. 104, No. 2, 1911, pp. 33, 47–48.

[12]Ellis, W. D., "Chadwick," *True's Automobile Yearbook,* No. 2, 1953, pp. 88–89, 132–133.

[13]Faustman, J. D., "Automatic Map Tracer for Land Navigation," *Electronics,* Vol. 17, No. 11, 1944, pp. 94–99.

[14]Rosen, D. A., Mammano, F. J., and Favout, R., "An Electronic Route Guidance System for Highway Vehicles," *IEEE Transactions on Vehicular Technology,* Vol. 19, 1970, pp. 143–152.

[15]French, R. L., and Lang, G. M., "Automatic Route Control System," *IEEE Transactions on Vehicular Technology,* Vol. 22, No. 2, 1973, pp. 35–41.

[16]Lezniak, T. W., Lewis, R. W., and McMillen, R. A., "A Dead-Reckoning/Map-Correlation System for Automatic Vehicle Tracking," *IEEE Transactions on Vehicular Technology,* Vol. 26, No. 1, 1977, pp. 47–60.

[17]Peters, T. J., "Automobile Navigation Using a Magnetic Flux-Gate Compass," *IEEE Transactions on Vehicular Technology,*" Vol. 35, No. 2, 1986, pp. 41–47.

[18]Ikeda, H., Kobayashi, Y., and Kawamura, S., "Sumitomo Electric's Navigation Systems for Private Automobiles," *Proceedings, VNIS '91—Vehicular Navigation & Information Systems Conference,* Vol. 1, Society of Automotive Engineers, Dearborn, MI, Oct. 20–23, 1991 (SAE Paper 912789), pp. 451–462.

[19]Silver, J., "The GBF/DIME System: Development, Design, and Use," Paper presented at the 1977 Joint Annual Meeting of American Society of Photogrammetry and American Congress on Surveying and Mapping, 1977.

[20]French, R. L., "Map Matching Origins, Approaches, and Applications," *Proceedings, Second International Symposium on Land Vehicle Navigation,* (Münster, Germany), July 4–7, 1989, 93–116.

[21]Karimi, H. A., private communication, Athabasca University, Edmonton, AB, Canada, March 16, 1993.

[22]French, R. L., "Mobile Communication," *Concise Encyclopedia of Traffic & Transportation Systems,* edited by M. Papageorgiou, Pergamon, New York, 1991, pp. 263–268.

[23]Shute, S., "RDS, The EBU Radio Data System," *International Broadcast Engineer,* May–July 1987.

[24]Honey, S. K., and Zavoli, W. B., "A Novel Approach to Automobile Navigation and Map Display," *Proceedings, NAV 85-Royal Institute of Navigation Conference on Land Navigation and Location for Mobile Applications,* Paper 27, York, England, Sept. 9–11, 1985.

[25]Buxton, J. L., Honey, S. K., Suchowerskyj, W. E., and Tempelhof, A., "The Travelpilot: A Second-Generation Automotive Navigation System," *IEEE Transactions on Vehicular Technology,* Vol. 40, No. 1, 1991, pp. 41–44.

[26]Ishikawa, K., Ogawa, M., Azuma, S., and Ito, T., "Map Navigation Software of the Electro-Multivision of the '91 Toyota Soarer," *Proceedings, VNIS '91—Vehicular Navigation & Information Systems Conference,* Vol. 1, Society of Automotive Engineers, Dearborn, MI, Oct. 20–23, 1991 (SAE Paper 912790), pp. 463–473.

[27]Umeda, Y., Morita, H., Azuma, S., and Itoh, T., "Development of the New Toyota Electro-Multivision," SAE Paper 920601, 1992.

[28]Ito, T., Azuma, S., and Sumiya, K., "Development of the New Navigation System—Voice Route Guidance," SAE Paper 930554, 1993.

[29]Krage, M. K., "The TravTek Driver Information System," *Proceedings, VNIS '91—Vehicular Navigation & Information Systems Conference,* Vol. 2, Society of Automotive Engineers, Dearborn, MI, Oct. 20–23, 1991 (SAE Paper 912820), pp. 739–748.

[30]Rillings, J. H., and Krage, M. K., "TravTek: An Operational Advanced Driver Information System," *Proceedings of the 1992 International Congress on Transportation Electronics,* Society of Automotive Engineers, Dearborn, MI, 1992 (SAE P-260), pp. 461–472.

[31]McLellan, J. F., Krakiwsky, E. J., and Huff, D. R., "Fleet Management Trials in Western Canada," *Proceedings, VNIS '91—Vehicular Navigation & Information Systems Conference,* Vol. 2, Society of Automotive Engineers, Dearborn, MI, Oct. 20–23, 1991 (SAE Paper 912826), pp. 797–806.

[32]McLellan, J. F., Krakiwsky, E. J., Schleppe, J. B., and Knapp, P. L., "The NavTrax™ Fleet Management System," *Proceedings, PLANS '92, IEEE Position, Location, and Navigation Symposium,* (Monterey, CA), IEEE, New York, March 25–27, 1992, pp. 509–515.

[33]French, R. L., "Automobile Navigation: Where is it Going?" *Navigation—Land, Sea, Air, & Space,* edited by M. Kayton, IEEE Press, New York, 1990, pp. 101–107.

[34]"IVHS Strategic Plan: Report to Congress," U.S. Department of Transportation, December 18, 1992.

Chapter 11

Marine Applications

Jim Sennott* and In-Soo Ahn†
Bradley University, Peoria, Illinois 61625
and
Dave Pietraszewski‡
United States Coast Guard Research and Development Center, Groton, Connecticut 06340

I. Marine Navigation Phases and Requirements

NAVIGATION can be defined as the process of planning, recording, and controlling the movement of a craft or vehicle from one place to another. It is a process that looks ahead in an effort to determine how a safe arrival can be secured. Physical sensors can only measure what "just" happened, at best. The navigation process, including sensor systems and associated human and electronic controls and monitor algorithms, must take this "historic" information and convert it into rudder and thrust commands that affect future events. How well this is achieved can be judged by various cost and safety measures, such as average vessel passage time, safe passage probability, and ship footprint control deviation.

In an effort to differentiate and quantify marine navigation safety requirements in the United States, the departments of Transportation and Defense have defined marine navigation in terms of "phases." The four phases of marine navigation defined in the Federal Radionavigation Plan[1] are ocean, coastal, harbor/harbor approach (HHA), and inland waterway. Each phase of marine navigation is distinguished by a clearly different set of performance requirements. These requirements are based on safety and environmental concerns and support the desire to minimize marine collisions, rammings, and groundings.

The current technological characteristics and policy constraints on the Global Positioning System standard positioning service (GPS SPS) will allow GPS to

This paper is a work of the U.S. Government and is not subject to copyright protection in the United States.

*Professor of Electrical and Computer Engineering.

†Associate Professor of Electrical and Computer Engineering.

‡Senior Navigation Scientist.

satisfy many of the ocean and coastal phase performance requirements. The same characteristics and constraints [particularly selective availability (SA)] make it unacceptable for HHA (and inland waterway) navigation. The major distinction is accuracy. The accuracy required for the ocean phase is 1800–3700 m (2 drms) and 460 meters (2 drms) for the coastal phase. The accuracy required for the HHA phase is 8–20 m (2 drms).

In the HHA phase, a vessel pilot needs accurate, frequent, and timely verification of the vessel's position. Deviation from the desired vessel track drives the pilot's decision-making process. This is quite unlike the ocean phase where position updates on the order of minutes are quite satisfactory. The need for frequent position verification places additional burdens on the radionavigation service provider. Momentary unexpected signal outages can significantly jeopardize the safety of vessels executing sensitive maneuvers. Akin to aircraft precision approach and landing guidance, an availability specification of 0.997 is currently stated for HHA, and integrity requirements will ultimately be established.

These specifications must be used with some caution. The stated accuracy and update specifications lack specificity with respect to dynamic conditions under which a stated level of sensor performance shall be provided. Moreover, no explicit allowance is made for pilot error contributed by personnel skill, pathway geometry, ship dynamics, and disturbances. The familiar concept of flight technical error employed in aircraft guidance has yet to have been systematically exploited in marine navigation design. Clearly, radionavigation requirements ultimately depend on vessel size, maneuvering activity, and the geographic constraints of the operating area. Ship pilotage may be unsafe, even with perfect navigation sensors, if maneuvers are complicated by tight channel tolerances, unexpected traffic, poor helm dynamics, and strong wind and current disturbances.

In developing the role of GPS and differential GPS (DGPS) in the marine environment, particular attention will be given to vessel footprint steering performance, and the interplay between sensor and ship models. Other related functions such as hazard warning, risk assessment, and on-line dynamics modeling, are also discussed. Before turning to the development of these GPS applications, some background on the experimentation that led to deployment of a standardized marine DGPS service in the United States is in order.

II. Marine DGPS Background

In the early 1980s, the Department of Transportation (DOT) began studying the potential civil use of GPS. DOT quickly realized that many potential applications would require higher levels of accuracy and integrity than SPS would be able to provide.[2]

The feasibility of providing DGPS for marine navigation with an absolute accuracy of 10 m (2 drms) was demonstrated in 1987. In 1989 the Montauk Point, New York radio beacon was temporarily converted to provide a DGPS test broadcast. Given the success of these 1989 tests, and the desire to evaluate DGPS in operational environments more thoroughly, a prototype DGPS service was established on August 15, 1990. The Montauk Point marine radiobeacon is now part of the U.S. Coast Guard Northeast testbed prototype service, providing accurate navigation from Cape Hatteras, North Carolina to Canada. The U.S.

Coast Guard has announced plans to cover most coastal areas of the United States by 1996.[3]

An essential part of the coastal service is a radio link for transmission of differential corrections. An approach that has been extensively investigated, and adopted, is MSK digital transmission of data over the existing network of low-frequency marine radiobeacons. The chosen modulation format minimizes interference with the radio direction finding function presently provided by these beacons. Furthermore, forward error detection and correction features are employed for reducing the impact of Gaussian and impulsive noise. This results in a highly reliable link, even beyond the normal rated range of radiobeacon service.[4,5]

Paralleling DOT DGPS sensor research, the radionavigation accuracy requirements for HHA and inland confined waterways were under study in the 80s. In 1987, an assessment of DGPS, differential Loran-C, and other candidates for the Great Lakes and St. Lawrence Seaway was carried out under U.S. Maritime Administration sponsorship.[6] Bradley University explored the interplay between sensor dynamics/noise characteristics and ship control performance in an automated steering environment. The steering characteristics of DGPS and other radionavigation sensor combinations were integrated with an augmented state navigation filter, and an optimal steering controller was developed.

In 1988 the Coast Guard Research and Development Center conducted human factors simulations with experienced pilots operating with simulated radio navigation sensors to validate the Federal Radionavigation Plan requirement for 8–20 m (2 drms) accuracy for harbor and harbor approach.[7] The study concluded that the 8–20 m requirement was appropriate for large vessels in restricted waterways. The study also concluded that radio aids to navigation with this level of accuracy would enhance traditional visual and radar navigation if implemented and used properly and would also enhance navigation under restricted visual conditions down to 0.25 nautical miles (n.mi.). However, the study noted that "additional understanding needs to be gained, particularly on the proper design and utilization of radio aid (RA) devices for negotiating turns under 0.0 n.mi. visibility conditions, before concluding that an RA system with 8–20 m (2 drms) accuracy can be used safely to support an all weather navigation system." Clearly, the usage of DGPS for steering in confined waterways will require careful integration with risk assessment and hazard warning functions that collect information from other sources such as radars, geographic data bases, and from shore side vessel traffic systems (VTS).

III. Global Positioning Systems-Assisted Steering, Risk Assessment, and Hazard Warning Systems

The overall model for GPS steering and hazard warning is shown in Fig. 1. The system is partitioned as follows:

1) Ship

Vessel steering hydrodynamics
Wind forces model
Water current model

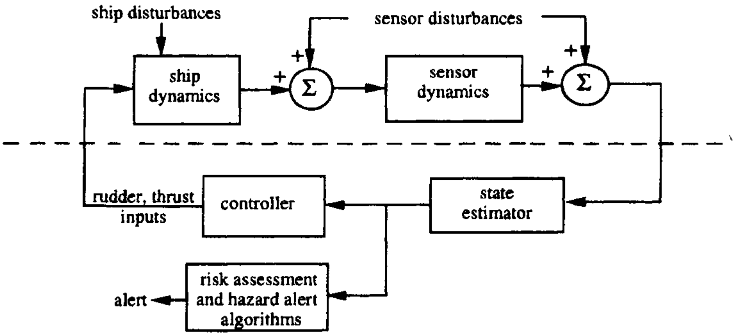
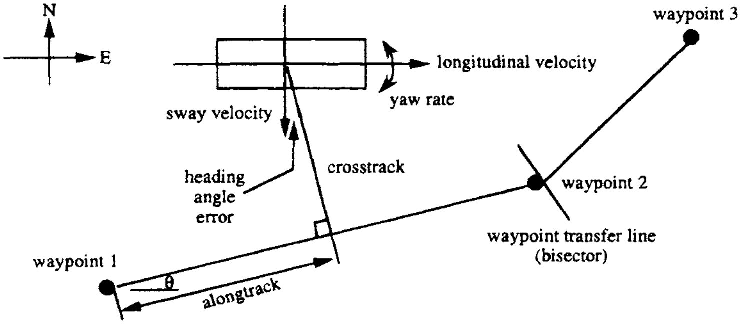


Fig. 1 Overall marine steering and hazard warning system.

2) Sensors
   Sensor lag dynamics
   Sensor noise disturbances
   User clock model
3) State estimation
   Nominal ship state variable model
   On-line vessel dynamics evaluator
   Navigation filter
4) Control
   Thrust control
   Rudder controller (human or automatic)
   Waypoint decision logic
5) Hazard warning and risk assessment
   Waypoint planning and risk evaluation
   Deviation and hazard warning algorithm

Figure 2 defines the coordinate systems and state variables that characterize the truth environment and form the basis for the navigation filter and rudder control design. Sway velocity, longitudinal velocity, and yaw rate are ship-referenced; heading angle error, crosstrack, and alongtrack position are waypoint referenced; current and wind components are east-north referenced.

1) Ship-fixed states
   x_{sway} = sway velocity
   x_{yaw} = yaw rate
   x_{vel} = longitudinal velocity
2) Waypoint referenced states
   x_{head} = heading angle error relative to waypoint segment
   x_{cross} = crosstrack error on a waypoint segment
   x_{along} = alongtrack position on a waypoint segment
3) East-north referenced states
   $x_{c,e}$ = crosstrack current bias
   $x_{c,n}$ = alongtrack current bias
   $x_{w,e}$ = crosstrack wind velocity component
   $x_{w,n}$ = alongtrack wind velocity component



ship dynamics states

sway velocity
longitudinal velocity
yaw rate

heading angle error
crosstrack
alongtrack

easting current
northing current
easting wind
northing wind

Fig. 2 Coordinate system and state definitions.

Ship dynamics are nonlinear and unstable. To implement the rudder control and navigation filter, a procedure for on-line linearization of hydrodynamics yielding stability derivatives as well as on-line linearization of the coupled state equations will be carried out.

The sensor portion of the system typically includes differential DGPS, Loran-C, ship-heading reference, waterspeed indicator, and wind anemometer. Careful attention must be given to the widely different sensor dynamics. When designing the navigation filter, state augmentation techniques can be applied to compensate for lags.

The rudder and thrust controller may consist of the human pilot with appropriate situation display, or automatic steering may be employed. As a human pilot navigates through a series of waypoints, a combination of factors are considered. If the potential for traffic conflict is low, the pilot generally applies rudder control so as to (subjectively) minimize the probability of violating the channel boundaries during the waypoint system passage. In the simulations presented later in this section, an optimal control law acts as "stand in" for the human pilot, with rudder control and navigation filter tuned to achieve the best performance from each sensor combination.

Although, ideally, system design would be based upon a safe passage probability (SPP) performance index accounting for waterway geometry, ship maneuverability, ship footprint size, and navigation sensor quality, the rigorous design of steering control under this criteria is a difficult unsolved problem. Therefore, more tractable design and evaluation criteria, described below, are employed.

The visual aid literature describes ship crosstrack error[8–10] and relative risk factor[7,11–14] performance indexes. Relative risk factor (RRF) assesses the minimum clearance between the vessel footprint and the channel boundary, at a specified alongtrack station. Accounting for the vessel footprint orientation in the channel, the RRF measure is closely related to the probability of grounding and collision at a specific channel location.

A more comprehensive measure is the channel clearance width distribution (CCWD).[6,15,16] The channel clearance width (CCW) process from which this distribution is derived is the collection of ship clearance envelopes swept out during many passages through the waypoint structure. The CCWD is simply the first-order cumulative distribution of ship CCWs, over all sample stations along the channel. This distribution will be developed below for a representative ship steering system driven by DGPS.

The last component shown in Fig. 1 supports risk assessment and hazard alert functions. These functions may be performed aboard ship and/or at a shoreside vessel traffic management site. For planning prior to vessel passage, a safe passage probability (SPP) performance measure could be computed for a given set of waypoints and vessel subsystems. The Global Positioning System can play a central role in supporting these calculations. Also, real-time supervisory alerts for fixed hazards and poor steering may be desired as a backup to human or autopilot steering and thrust control. One such function is based upon a running computation of latest safe alarm time (LSAT) for known hazards along the waterway.

IV. Vessel and Sensor Modeling

A. Vessel Dynamics Model

The three body-fixed ship states, sway velocity, yaw rate, and longitudinal velocity, are described by nonlinear coupled differential equations. Just as in aerodynamic flight, hydrodynamic stability derivatives are employed to express body-fixed dynamics in a linear fashion. A set of stability derivatives is valid in the region about which the system is linearized. These coefficients will vary depending upon longitudinal and sway velocities and yaw rate, and on ship draft and waterway bottom clearance. The equations governing body-fixed states are given below. The two control variables are rudder angle and propeller thrust.

$$\begin{aligned}\dot{x}_{\text{sway}} = {} & D_{s,s}x_{\text{sway}} + D_{s,y}x_{\text{yaw}} + D_{s,bw}\,\{[x_{w,n}\cos\theta + x_{w,e}\sin\theta]\sin x_{\text{head}} \\ & + x_{w,n}\sin\theta + x_{w,e}\cos\theta]\cos x_{\text{head}}\} + D_{s,\delta}\delta_{\text{rudder}}\end{aligned}$$

$$\begin{aligned}\dot{x}_{\text{yaw}} = {} & D_{y,s}x_{\text{sway}} + D_{y,y}x_{\text{yaw}} + D_{y,bw}\,\{[x_{w,n}\cos\theta + x_{w,e}\sin\theta]\sin x_{\text{head}} \\ & + x_{w,n}\sin\theta + x_{w,e}\cos\theta]\cos x_{\text{head}}\} + D_{y,\delta}\delta_{\text{rudder}}\end{aligned}$$

$$\begin{aligned}\dot{x}_{\text{vel}} = {} & D_{v,v}x_{\text{vel}} + D_{v,lw}\,\{[x_{w,n}\cos\theta + x_{w,e}\sin\theta]\cos x_{\text{head}} \\ & + [x_{w,n}\sin\theta + x_{w,e}\cos\theta]\sin x_{\text{head}}\} + D_{\text{v,thrust}}\delta_{\text{thrust}}\end{aligned}$$

where D coefficients, defined below, are the hydrodynamic and wind force coefficients of the ship, and θ is the waypoint leg heading. Linearized dynamics provide information about the stability of the above nonlinear dynamics. In linearized dynamics, in states x_{sway} and x_{yaw} there is one positive eigenvalue. Hence, the rudder controller design must provide ship stability.

Table 1 Ship parameters

Length of ship	305 m
Width of ship	38 m
Hull type	Tanker
Speed	5.14 m/s (10 k)
Dynamics coefficients	
$D_{s,s} = -.0145$	
$D_{s,y} = -2.477$	
$D_{s,\delta} = -.01515$	
$D_{y,s} = -.00029$	
$D_{y,y} = -.0897$	
$D_{y,\delta} = .000392$	

Stability derivatives are obtained by taking first-order derivatives of nonlinear hydrodynamic equations at the nominal operating point. In simulation study, these derivatives are later matched to follow actual ship dynamics. For an example, see Table 1.

$D_{s,s}$ = sway dynamics coefficient
$D_{s,y}$ = yaw-to-sway coupling coefficient
$D_{s,\delta}$ = rudder-to-sway coupling coefficient
$D_{y,s}$ = sway-to-yaw coupling coefficient
$D_{y,y}$ = yaw dynamics coefficient
$D_{y,\delta}$ = rudder-to-yaw coupling coefficient
$D_{v,v}$ = longitudinal velocity dynamics coefficient
$D_{v,\text{thrust}}$ = thrust-to-longitudinal velocity coupling coefficient
$D_{s,bw}$ = beam wind-to-sway coupling coefficient
$D_{y,bw}$ = beam wind-to-yaw coupling coefficient
$D_{v,lw}$ = longitudinal wind-to-longitudinal velocity coupling coefficient

In addition to the hydrodynamic effects, wind forces play a very important part in the vessel motion. In the preceding list, coupling of winds into dynamics is treated as follows. Neglecting nonhomogeneous effects, the true wind vector is first resolved into the body-fixed coordinate frame. Then the relative beam and bow wind components are determined. It is assumed that these components induce torque and force terms linear in wind component magnitude. For example, the beam wind component generates beam force and torque values proportional to wind component magnitude.

The next three state equations describe ship position relative to the waypoint system. The coupling of longitudinal and sway velocity into alongtrack and crosstrack equations is nonlinear because of the rotation between waypoint and body-fixed frames.

$$\dot{x}_{\text{head}} = x_{\text{yaw}}$$

$$\dot{x}_{\text{cross}} = x_{\text{sway}} \cos x_{\text{head}} + x_{c,e} \cos \theta - x_{c,n} \sin \theta + x_{\text{vel}} \sin x_{\text{head}}$$

$$\dot{x}_{\text{along}} = -x_{\text{sway}} \sin x_{\text{head}} + x_{c,e} \sin \theta + x_{c,n} \cos \theta + x_{\text{vel}} \cos x_{\text{head}}$$

The current and wind states, in the easting-northing frame, are expressed with four linear state equations

$$\begin{aligned}\dot{x}_{c,e} &= -\alpha_c x_{c,e} + w_{c,e}\\ \dot{x}_{c,n} &= -\alpha_c x_{c,n} + w_{c,n}\\ \dot{x}_{w,e} &= -\alpha_w x_{w,e} + w_{w,e}\\ \dot{x}_{w,n} &= -\alpha_w x_{w,n} + w_{w,n}\end{aligned}$$

where the forcing terms are chosen to model the random wind and water current terms.

A linear state space representation for the ship must be developed from these nonlinear state equations prior to obtaining the rudder controller and navigation filter. The state matrix coefficients of the continuous-time incremental linear model are derived by taking partial derivatives with respect to states and evaluating at an assumed operating point. Then the incremental continuous-time linear model may be written in the usual form as follows

$$\Delta\dot{\boldsymbol{x}}^{\text{ship}} = A_{\text{ship}}\Delta\boldsymbol{x}^{\text{ship}} + \boldsymbol{Bu}$$

where

$$\Delta\boldsymbol{x}^{\text{ship}} = (\Delta x_{\text{sway}}\Delta x_{\text{yaw}}\Delta x_{\text{vel}}\Delta x_{\text{head}}, \Delta x_{\text{cross}}\Delta x_{\text{along}}\Delta x_{c,e}\Delta x_{c,n}\Delta x_{w,e}\Delta x_{w,n})^t$$

and

$$\boldsymbol{u} = (\delta_{\text{rudder}}\ \delta_{\text{thrust}})^t$$

with superscript t denoting transpose.

B. Standardized Sensor Model

The fundamental quantities observed in GPS or Loran-C systems are pseudorange and Doppler. These are corrupted by noise and dynamical errors. Noise errors are those observed at sensor outputs in the absence of any vehicle motion. It is often assumed that these are uncorrelated in time. Upon closer examination such jitter errors are found to be colored, reflecting both receiver front-end noise and signal tracking loop properties. Dynamical errors in pseudorange and Doppler result from the filtering actions of signal tracking loops. In critical vessel control problems these errors can be a very significant part of the overall error budget. Indeed, mariners using Loran-C are accustomed to lags between actual and reported position, compensating to a degree in their vessel pilotage. To compare different sensor combinations clearly for the precise navigation problems of interest, a linear equivalent sensor dynamics model is used to portray the tracking loop for each sensor channel.

Consider a generic sensor that tracks pseudorange. The navigator observes pseudorange via the navigation sensor hardware, in GPS from the delay-lock loop, and in Loran-C from the third cycle zero-crossing tracker. Both noise and dynamic lag terms can distort true pseudorange. Therefore, it is desirable to include these effects in the system design and evaluation. Let R_i be the pseudorange to the ith sensor. Then, the following equation results

$$R_i(t) = \sqrt{[e_{\text{ship}}(t) - e_{\text{sensor},(t)}]^2 + [n_{\text{ship}}\,(t) - n_{\text{sensor},i}(t)]^2} + x_{\text{clock}}(t)$$

where $e_{\text{ship}}(t)$ and $n_{\text{ship}}(t)$ denote the ship's easting and northing positions at time t, respectively. Similarly, $e_{\text{sensor},i}(t)$ and $n_{\text{sensor},i}(t)$ denote the ith sensor's easting and northing positions, respectively. The $e_{\text{ship}}(t)$ and $n_{\text{ship}}(t)$ variable are nonlinear, through geometry, in ship along-track and cross-track states. The state $x_{\text{clock}}(t)$ is the offset between user and system time. Let $\boldsymbol{R}$ be the collection of true pseudorange, and Doppler values for the following sensor set

$$\boldsymbol{R} = (R_1\ R_2 \cdots R_{n-1}\ R_n)^t$$

A linear state equation representation portrays demodulator loops appropriate for both GPS and Loran-C tracking channels. Sensor states are the quantities actually seen as navigation filter inputs. The dimension of A_{sensor} is chosen to match the tracking loop order and number of channels employed. The noise power spectral density of w^{sensor} models equivalent atmospheric and/or receiver front-end sources. The forcing vector $\boldsymbol{R}$ is as defined above.

$$\dot{\boldsymbol{x}}^{\text{sensor}} = A_{\text{sensor}}\boldsymbol{x}^{\text{sensor}} + B_{\text{sensor}}\boldsymbol{R} + \boldsymbol{w}^{\text{sensor}}$$

C. Combined Ship and Sensor Model

To form a complete model the sensor state equations are combined with the previously derived ship state equations. To do so, the state nonlinearity in $\boldsymbol{R}$ is linearized about the nominal ship state solution. Periodically the nominal ship solution is updated and the incremental ship state is reset to zero

$$\boldsymbol{x}_{\text{total,ship}} = \boldsymbol{x}_{\text{nominal,ship}} + \Delta\boldsymbol{x}^{\text{ship}}$$

where $\Delta\boldsymbol{x}^{\text{ship}}$ is the incremental solution defined above, and the nominal dynamics are treated as an unforced system with nonzero initial conditions.

Then, the sensor forcing terms in vector $\boldsymbol{R}$ are expanded about the nominal solution, the incremental term providing the desired connection between the ship dynamics and sensor dynamics. In continuous time, the combined sensor and incremental ship state equations are given by the following

$$\begin{bmatrix} \Delta\dot{\boldsymbol{x}}^{\text{ship}} \\ \dot{\boldsymbol{x}}^{\text{sensor}} \end{bmatrix} = \begin{bmatrix} A_{\text{ship}} & 0 \\ A_{\text{ship, sensor}} & A_{\text{sensor}} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{x}^{\text{ship}} \\ \boldsymbol{x}^{\text{sensor}} \end{bmatrix} + \begin{bmatrix} B_{\text{ship}} & 0 \\ 0 & B_{\text{sensor}} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{\text{ship}} \\ \boldsymbol{u}^{\text{sensor}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{w}^{\text{ship}} \\ \boldsymbol{w}^{\text{sensor}} \end{bmatrix}$$

where

$$\boldsymbol{x}^{\text{sensor}} = \text{navigation filter input}$$

with

$$\boldsymbol{u}^{\text{ship}} = [\delta_{\text{rudder}}\ \delta_{\text{thrust}}]^t$$

and

$$\boldsymbol{u}^{\text{sensor}} = [R_1\ R_2\ R_3 \cdots R_n]^t_{\text{nominal}}$$

From the continuous-time system the usual discrete-time linear system is obtained (Ref. 17)

$$\Delta \boldsymbol{x}(k+1) = \Phi \Delta \boldsymbol{x}(k) + \Gamma \boldsymbol{u}(k)$$

where the state transition matrix Φ is given by

$$\Phi = e^{A\Delta t}$$

and

$$\Gamma = \int_0^{\Delta t} e^{A\xi} B \mathrm{d}\xi$$

In discrete time the complete model is given by the following

$$\begin{bmatrix} \Delta \boldsymbol{x}^{\text{ship}}(k+1) \\ \Delta \boldsymbol{x}^{\text{sensor}}(k+1) \end{bmatrix}$$

$$= \begin{bmatrix} \Phi_{\text{ship}} & 0 \\ \Phi_{\text{ship,sensor}} & \Phi_{\text{sensor}} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{x}^{\text{ship}}(k) \\ \Delta \boldsymbol{x}^{\text{sensor}}(k) \end{bmatrix} + \begin{bmatrix} \Gamma_{\text{ship}} & 0 \\ 0 & \Gamma_{\text{sensor}} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}^{\text{ship}} \\ \boldsymbol{u}^{\text{sensor}} \end{bmatrix} + \text{noise}$$

It is important to note the ship state transition matrix is not static. While the ship is underway its hydrodynamics vary, because of changes in channel bottom and side clearance and large variations in vessel side-slip, yaw rate, and longitudinal velocity. Therefore, it is desirable to monitor and update the hydrodynamic coefficients by on-line identification techniques.[18–20] GPS-derived position, velocity, and attitude data are an excellent basis for this on-line hydrodynamic modeling. In any event, the ship state transition matrix is impacted by rotations between body-fixed and waypoint-fixed systems; therefore, updating of the above is performed whenever a significant change in ship heading is detected. In so doing, the nominal ship solution is set to the present estimate of total ship state, and the ship incremental state estimate is reset to zero. These updates impact both the navigation filter and rudder controller portions of the system.

V. Waypoint Steering Functions

Vessel steering in confined waterways is at present conducted manually, with the great majority of steering performance studies focused upon human factors simulations and shipboard observations of pilot performance. However, DGPS offers the possibility of a transition to a more automated system. In anticipation of such systems, and with an additional goal of understanding how well human pilots might perform with DGPS under ideal conditions, a simulator was developed encompassing the above ship and sensor models.[6,15,16] In this work, an automatic steering system acts as stand-in for the human pilot, a simplification that sets aside the admittedly complex issue of pilot display configuration and man-machine interface. The discrete-time ship and sensor model derived above provides a foundation for the autopilot and navigation filter design. The most promising configurations can later be evaluated by full-fledged human factors simulation.

A. Filter and Controller Design

The ship navigation filter developed for this study integrates a variety of radionavigation sensors, each properly characterized in terms of equivalent noise sources and signal tracking loop dynamics. In the filter, a reduced-order ship model with constant thrust is assumed. Five ship states are included: two in the body-fixed frame, sway velocity and yaw rate, and three in the local waypoint coordinate system frame, alongtrack position, crosstrack error and heading. Two water current velocity states, easting and northing, are also estimated. Random wind disturbances are introduced on sway and yaw rate states. In the navigator, these seven states portray the ship motion more accurately than a simpler easting–northing or Earth-centered, Earth-fixed (ECEF) based filter. This filter model entails nonlinearities in the system dynamics matrix, requiring periodic linearization about the estimated ship trajectory.

Within the navigation filter, seven additional states are associated with the sensor model. As shown previously, these may be modeled together with five ship incremental states as one larger linear system. Five of the seven sensor states portray sensor smoothing lags, as encountered in the pseudorange and heading sensor elements of the navigation system. The two remaining sensor states model the navigation receiver clock bias and bias rate internal to the DGPS and differential Loran-C receiver.

Navigation filter estimates of all ship states are passed on to the rudder controller, whose job is to position and orient the ship footprint relative to the specified channel boundaries. When steering large ships in confined waterways, attention must be given to both crosstrack steering and yaw errors. Specifically, control seeks to minimize the CCW, defined as the minimum channel width that will clear the ship footprint. To be viable, such a control must tolerate random disturbances of current, wind, and model mismatch, making best use of the complete navigation filter state vector. In turning maneuvers, optimality under the CCW criteria is an unsolved problem. However, in straight track-keeping segments, CCW may be formulated as a stochastic regulator problem. To handle both straight and turning segments, the implemented steering algorithm consists of 1) an "inner" steering regulator designed to maintain tight control of ship cross-track and yaw error on each leg of the waypoint structure, and 2) an "outer" decision algorithm for transitioning between waypoint segments.

The inner control is a stochastic regulator whose goal is to drive waypoint-referenced error states to zero. The following quadratic performance index is utilized

$$
\begin{aligned}
J_i &= \frac{1}{2}\sum_{k=i}^{N-1} \{q_h \Delta x_{\text{head}}^2 + q_c \Delta x_{\text{cross}}^2 + r_{11}\delta_{\text{rudder}}^2 + r_{22}\delta_{\text{thrust}}^2\} \\
&= \frac{1}{2}\sum_{k=i}^{N-1} \{\Delta \boldsymbol{x}'\,(k)\; Q\; \Delta \boldsymbol{x}(k) + \boldsymbol{u}'\,(k)\; r\; \boldsymbol{u}(k)\}
\end{aligned}
$$

The weighting matrix Q consists of diagonal elements q_h and q_c, and other arbitrarily small positive elements along the diagonal to make Q positive semidefinite. This choice of Q implies emphasis on heading angle error and crosstrack error, possibly with some weighting of sway velocity and yaw rate. The matrix r for the control input is positive definite with r_{11} for the rudder control and r_{22} for the thrust control. Using the above Q and r matrices the optimal control u is obtained.[21]

B. Sensor/Ship Bandwidth Ratio and Straight-Course Steering Performance

A very simple straight track-keeping scenario is first explored. This gives an insight into the interaction of GPS and LORAN signal-tracking parameters with vessel dynamics parameters. Consider for the moment, lateral motion control and minimization of crosstrack error. The ship bandwidth on this axis was adjusted to a typical value for a medium-sized ocean vessel, 0.05 rad/s. Three different random vessel disturbance levels were considered, resulting in open-loop crosstrack standard deviations of 50, 15, and 5 m, respectively.

In this simplified model, an equivalent one-dimensional sensor was employed for estimating crosstrack displacement. First-order sensor dynamics were modeled, with sensor state process noise adjusted to maintain a constant crosstrack sensor error variance for each bandwidth setting considered. The navigation filter included a sensor lag state properly matched to this model. Then, an optimum linear quadratic Gaussian (LQG) controller was derived and the steady state crosstrack performance index for the combined filter and regulator was developed analytically.

The achieved crosstrack steering performance was strongly influenced by the sensor/vessel bandwidth ratio. Very poor control was obtained with a sensor bandwidth narrower than the vessel bandwidth. To underscore this effect, the steering error statistics were plotted in terms of the sensor/vessel bandwidth ratio, shown in Fig 3. In all cases, the sensor output standard deviation was held to

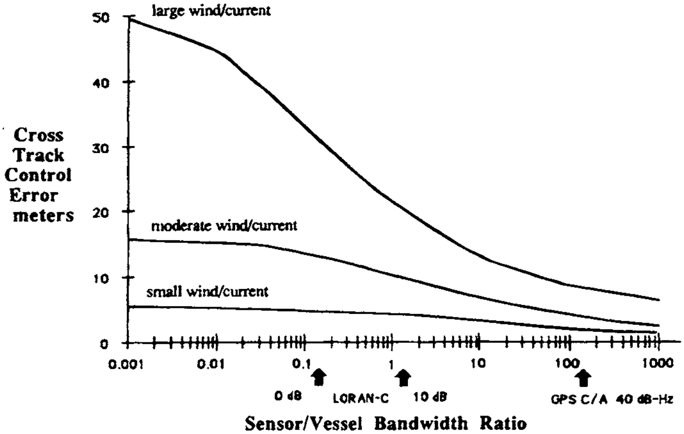


Fig. 3 Vessel steering performance vs bandwidth ratio.

5 m. At very small sensor/vessel bandwidth ratios, the controller is essentially operating without a sensor, and the crosstrack errors are the open-loop values of 50, 15, and 5 m. At unity bandwidth ratio, errors decreased to 21, 10, and 4.5 m, the largest improvement for the high wind/current disturbance scenario.

The above generic results offer insight into comparative DGPS and Loran-C performance. For a 20-kHz signal-to-noise ratio (S/N) of + 0 dB, a typical Loran-C receiver operates with a bandwidth of 0.007 rad/s. At +10 dB S/N the receiver loop bandwidth may be widened to 0.07 rad/s. Loran-C falls between 0.14 and 1.4 on the sensor/vessel bandwidth axis of Fig. 3. For a coherent delay-lock GPS tracking loop operating at a carrier-to-noise ratio of 40 dBHz and with the assumed 5 m standard deviation, the bandwidth is considerably broader, approximately 6 rad/s. Global positioning systems C/A reception at 40 dBHz carrier-to-noise ratio (C/N) achieves a bandwidth ratio of 125, permitting considerably better vessel control performance.

For the large wind/current example, the control performances for the different sensors fall between 33 and 9 m, crosstrack error. Although both Loran-C and GPS receivers can be designed to have identical noise jitter, five meters in the above analysis, the much wider bandwidth of GPS affords far better vessel control. Global positioning systems Doppler observables, or carrier-smoothed pseudoranges, would further enhance this performance. Human pilots tend to compensate for lags in guidance sensors. This subjective behavior has been formalized in the above by inclusion of lag states in the navigation filter design. We now return to the complete waypoint simulation.

C. Comparative Footprint Channel Clearance Width Distributions

Prior to closed loop operation with the sensor system and navigation filter, ship hydrodynamic coefficients were calibrated against published at-sea test data. The general vessel parameters considered are as shown in Table 1. An important test of ship dynamics is the 20/20 Z maneuver test.[8] From a straight line path with rudder amidships, the helm (rudder command) is deflected 20 deg to the right. When the heading changes 20 deg to the right of the initial heading, the helm is reversed to 20 deg left. When the heading changes to 20 deg left of the initial heading, the helm is reversed to 20 deg right. Beginning with available hydrodynamic coefficients,[19] coefficients were trimmed to give 20/20 Z maneuver dynamics similar to reported at-sea tests.[8] In an operational system, these model coefficients could be continuously refined by on-line identification techniques.

The basic test scenario consisted of a strait approach segment of 1500 m, a 35-deg/course change to port, and another 2500 m of course keeping. A diminishing southwesterly current of 1.5 k was applied early in each run, and to simulate the effect of wind and unmodeled hydrodynamic forces, random disturbances were introduced on the ship sway velocity and yaw rate. The higher of the two disturbance levels was selected to model storm conditions likely to be encountered only infrequently in river and harbor areas. The low-disturbance scenario is more typical of everyday piloting.

Even with perfect knowledge of ship position, velocity, heading, and yaw rate states, some deviation from the desired waypoint trajectory is unavoidable. Thus, before examining GPS and other sensor systems, it was important to determine steering errors contributed by ship controllability factors and wind and water disturbances. To this end, the optimal ship controller was driven with a perfect navigation state vector under a variety of disturbance conditions, and baseline values for sway and yaw rate disturbances were established. Also, the "outer" decision control for course leg switching points was optimized to obtain best CCW performance.

Figure 4a shows typical clearance widths swept out by the ship footprint for two disturbance levels in the region just after the turn. The graph ordinate is accumulated alongtrack position, and the abscissa is the minimum channel width needed to just clear the extremity of the ship. With small ship disturbance, the turn approach and postturn recovery clearance values are nearly the same as the ship half width, 19 m. Not shown is the turn region, where the right clearance width must be increased to 150 m and the left clearance width must be increased to 80 m to avoid boundary contact. The high disturbance CCW values superimposed on the plot show substantial increases in needed channel width for safe passage.

To further quantifying this behavior, the cumulative probability distribution for CCW samples at alongtrack points was estimated by simulation. Each distribution is the result of several runs and thousands of data points. Because CCW distributions are likely to differ in approach, turn, and recovery regions, they should be computed separately for each region. Figure 4b compares the CCW distributions in the postturn region, for both low- and high-disturbance conditions. For the low-disturbance runs, 95% of the clearance width samples are under 30 m. For the high-disturbance group, the 95% clearance half width is 48 m. An analytical derivation of these distributions is under development, with the ultimate goal of determining safe passage probability.

As discussed earlier in this book, GPS user and reference station equipment can operate in either code mode or integrated Doppler mode, with the most accurate results obtained when both sites are in the integrated Doppler mode. Delta range is useful in estimating velocity states during the required "warm-up" interval of the integrated Doppler mode. In the following simulations, perfect reference station corrections are assumed, together with a shipboard DGPS receiver operating with unsmoothed code pseudorange. In this GPS configuration, the sensor error budget is dominated by code-tracking loop noise, multipath, and tracking loop dynamical errors. Commensurate with a received C/N of 40 dBHz, and a tracking loop jitter of 5 m, the simulated code tracking loop time constant was set at 1.0 s. The navigation filter also processes a heading sensor good to 1 deg, 1-sigma. All data are sampled at a 1 Hz rate.

Fig. 5a shows typical GPS CCW plots obtained under low- and high-disturbance conditions, for the postturn region. At the low-disturbance level, CCW is very close to that obtained with perfect ship state knowledge. It is instructive to compare CCW cumulative distributions against the perfect state knowledge case. These are shown in Fig. 5b. The DGPS CCW distribution closely tracks the perfect-state-knowledge case. At the 95% level, an increase of about 2 m in CCW is contributed by the DGPS sensor and heading reference system. The sensor system contribution is greater with large disturbance, about 16 m. This

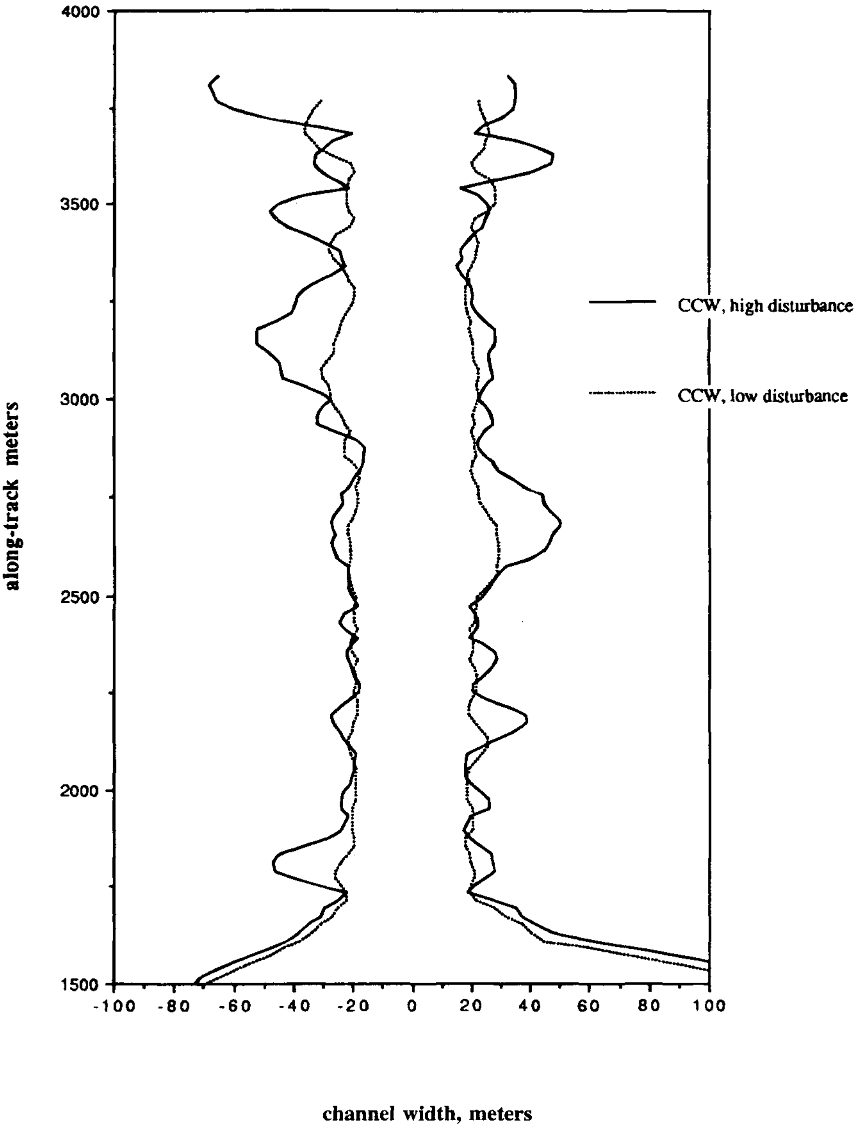


Fig. 4a Channel clearance width with perfect navigation sensors.

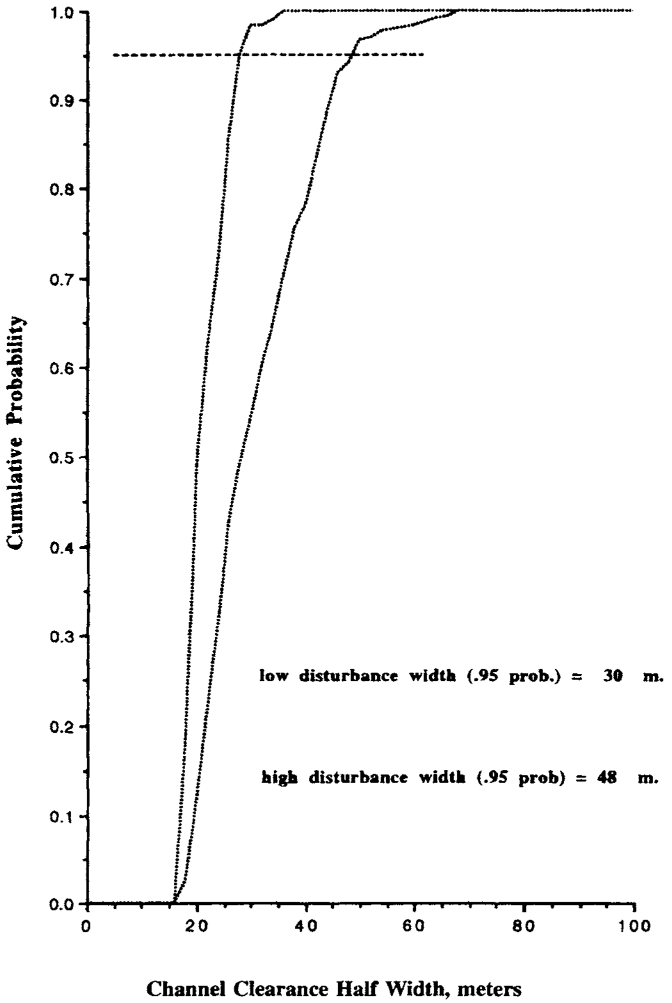


Fig. 4b Channel clearance width distributions, perfect nav.

could be significantly improved with delta range velocity or smoothed pseudorange observables.

Loran-C may be operated differentially to achieve improved accuracy. Calibration is problematical, however, given local grid warp conditions in river and harbor areas. Furthermore, from the earlier discussion, signal tracking loop dynamics will significantly degrade steering performance. To understand better its fundamental performance limits, a perfectly calibrated Loran grid was assumed. Then, a simulated tracking loop jitter of 5 m, identical to the GPS pseudorange jitter above, was introduced. For an assumed 20-kHz S/N of +10 dB this corresponds to a tracking loop time constant of about 14 s. Clock and heading reference parameters and rudder data quality were the same as for GPS.

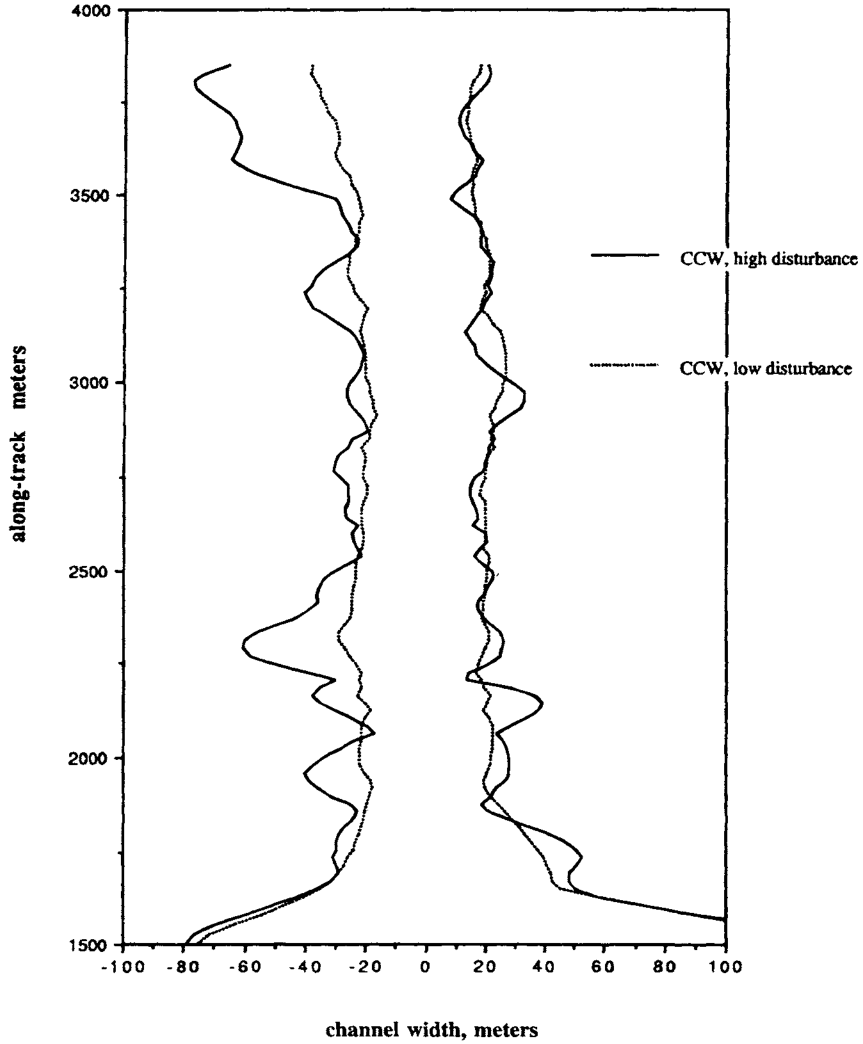


Fig. 5a Channel clearance width with differential GPS.

Figure 6a shows typical CCW results for high- and low-disturbance conditions in the turn recovery region. The plot exhibits a good deal of steering drift, indicative of poorer velocity estimates within the navigation filter, and a consequent loosening of the ship control loop. Turning to the cumulative distributions for channel clearance (Fig. 6b), a substantial loss in performance over the perfect-nav case is observed. At the 95% level, an increase of about 37 m in CCW is observed. This is with a five meter pseudorange jitter, and with sensor lag augmentation in the nav filter.

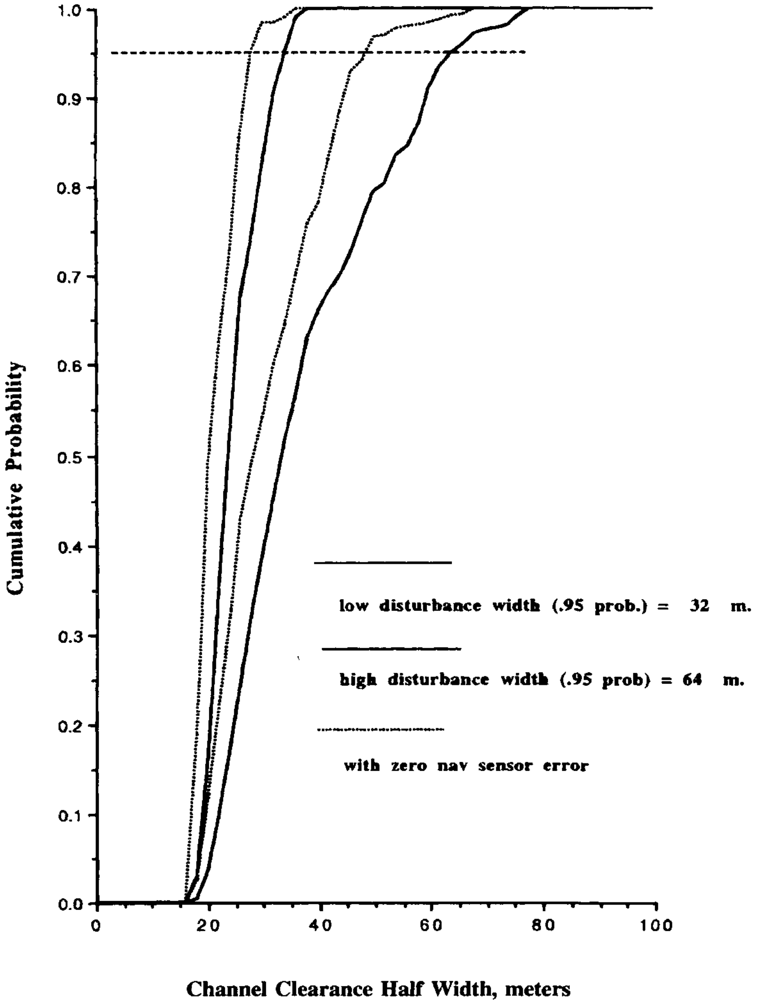


Fig. 5b Channel clearance width distributions, differential GPS.

In summary, although tracking loop output standard deviation characteristics and sensor sample rates were identical for both DGPS and Loran-C tests, implying the same static accuracy for both sensors, achieved control performances differed markedly. As these results clearly show, specification of receiver jitter and update rate alone is insufficient for predicting ultimate closed-loop steering performance. Sufficiency requires characterization of both sensor dynamics and sensor noise sources. In the aviation environment, this issue is of even greater significance.

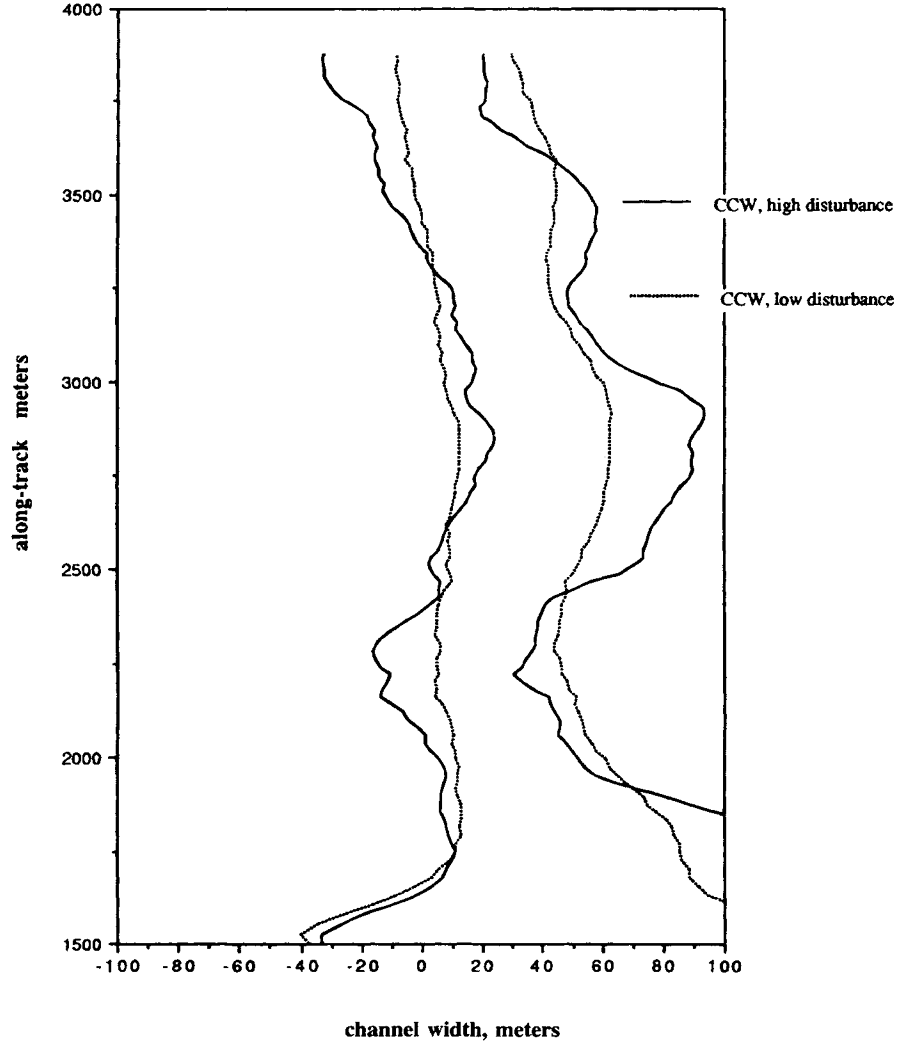


Fig. 6a Channel clearance width with Loran-C.

VI. Hazard Warning and Risk Assessment Functions

A. Risk Assessment

A preview of risk in advance of vessel passage through a waterway system, considering waterway geometry, anticipated steering quality, and navigation system anomalies can be performed aboard ship or at the shore-based vessel traffic system. A well-designed graphics interface would enhance the usefulness of this tool for testing "what if" scenarios. A safe passage probability figure of merit could be used to rank alternative waypoint steering sequences prior to passage.

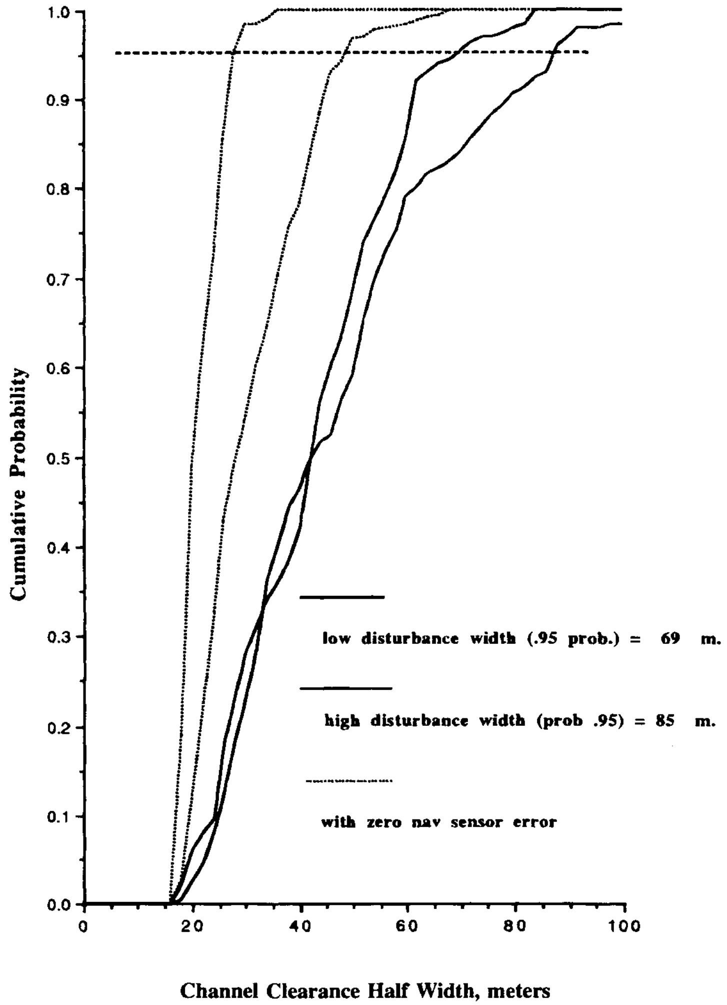


Fig. 6b Channel clearance width distributions, Loran-C.

The computation of risk depends upon knowledge of ship-specific hydrodynamic coefficients obtained over a history of runs in the same, or similar, confined waterway geometries. Differential GPS could be the basis for estimation of these ship coefficients, which would be entered into a closed-loop steering simulation similar to that discussed in Sec. V. Operators would input desired waypoints either by keyboard or by light-pen on a map display. The following computations would then be carried out: 1) optimization of waypoint segment switch points; 2) generation of the disturbance-free nominal-ship/nominal-sensor trajectory; 3) introduction of the ship disturbance model; 4) introduction of the sensor disturbance model; 5) computation of first- and second-order CCWDs; and 6) evaluation of SPP.

Computations can be computed for a specified waypoint sequence in several seconds with today's processors. In addition to statistical data, graphical outputs overlaid on the electronic chart would include points of waypoint turn initiation and the resulting ship footprint sweeps.

B. Hazard Warning

In contrast to the planning aspect of risk assessment, hazard warning addresses the immediate threat from nearby traffic and fixed hazards. Warning of conflict with other vessels requires continuous updates of relative position and the exchange of maneuver information confirming rules-of-road procedures. Warning of fixed hazards requires integration of the geographic database with ship and environmental states. GPS plays a vital role here.

Present-day vessel collision avoidance systems acquire relative positions from short-range marine radar. Automatic exchange of GPS position, velocity, and identity via a dedicated vhf channel would greatly improve the quality of this collision avoidance data. The beacon device would consist of a modified vhf transceiver connected to an existing GPS receiver. A similar GPS position reporting beacon could be affixed by authorities to new or critical hazards. At a very low incremental cost, smaller vessels that frequent busy commercial shipping lanes could also participate by equipping themselves with a GPS reporting beacon.

State estimates derived from GPS can be used to support a warning algorithm with the following properties. First, it does not generate an alarm (false) when the hazard miss distance would be acceptable in the absence of an alarm-induced maneuver. Second, it does generates an alarm (successful) in sufficient time for avoidance when the miss distance would otherwise be unacceptable in the absence of an alarm-induced maneuver.

The trade between successful and false alarms has been evaluated for an algorithm that estimates the Latest Safe Alarm Time (LSAT). Making use of the present state vector estimate, the vehicle maneuver model, and current rudder activity if available, the algorithm continuously probes for the latest time a maneuver may be undertaken to avoid the given hazard boundaries successfully. It has been shown that GPS velocity data can greatly improved the trade between the false and successful alarms.[22]

VII. Summary

The phase-in of more automated control and warning systems must include a period where their information is used in an advisory capacity. This provides an opportunity to identify and evaluate the remaining problems with the implementation without risking the safety of the crew and vessel involved. The phases of navigation clearly represent distinct levels of risk. An integrated approach using a technology such as DGPS combined with other sensors and the actual maneuvering characteristics of the vessel will be needed. In conclusion, GPS/DGPS means much more to marine transportation than "where the vessel is located." The investigation of other potential benefits and associated technologies as yet unforeseen will advance transportation safety and economic benefits in the future.

References

[1]"Federal Radionavigation Plan," U.S. Department of Transportation.

[2]Hartberger, A., "Introduction to the U.S. Coast Guard Differential GPS Program," *Proceedings, IEEE PLANS '92, IEEE Position, Location, and Navigation Symposium,* (Monterey, CA), IEEE, New York, March 25–27, 1992.

[3]Spalding, J., and Krammes, S., "Impact of USCG Differential GPS Service on User Equipment Technology," *Proceedings GPS-92,* Sept. 1992, pp. 573–578.

[4]Enge, P. K., Ruane, M., and Sheynblatt, L., "Marine Radiobeacons for the Broadcast of Differential GPS Data," *Proceedings of IEEE PLANS,* IEEE, New York, Nov. 1986, pp. 368–376.

[5]Enge, P. K., Kalafus, R. M., and Ruane, M. F., "Differential Operation of the Global Positioning System," *IEEE Communications Magazine,* July 1988, pp. 48–60.

[6]Sennott, J. W., and Ahn, I. S., "Assessment of Candidate Navigation Systems for Great Lakes St. Lawrence Seaway," Rept. of U.S. Department of Transportation Contract DTRS-57-85-C-0090, TTD 3, March 1988.

[7]Gynther, J. W., and Smith, M. W., "Radio Aids to Navigation Requirements: The 1988 Simulator Experiment," U.S. Coast Guard Rept. CG-D-08-90, Dec. 1989.

[8]Bertsche, W. R., Atkins, D. A., and Smith, M. W., "Aids to Navigation Principal Findings Report on the Ship Variables Experiment: The Effect of Ship Characteristics and Related Variables on Piloting Performance," U.S. Coast Guard Rept. CG-D-55-81, Nov. 1981.

[9]Bertsche, W. R., and Smith, M. W., "Aids to Navigation Principal Findings on the CAORF Experiment—The Performance of Visual Aids to Navigation as Evaluated by Simulation," U.S. Coast Guard Rept. CG-D-51-81, Feb. 1981.

[10]Bertsche, W. R., Mirino, K. L., and Smith, M. W., "Aids to Navigation Principal Findings Report: The Effect of One-Sided Channel Marking and Related Conditions on Piloting Performance," U.S. Coast Guard Rept. CG-D-56-81, Dec. 1981.

[11]Cooper, R. B., Cook R. C., and Marino, K. L., "At-Sea Data Collection for the Validation of Piloting Simulation," U.S. Coast Guard Rept. CG-D-60-81, Dec. 1981.

[12]Marino, K. L., Smith, M. W., and Moynehan, J. D., "Aids to Navigation SRA Supplemental Experiment Principal Findings: Performance of Short Range Aids Under Varied Shiphandling Conditions," U.S. Coast Guard Rept. CG-D-03-84, Sept. 1984.

[13]Moynehan, J. D., and Smith, M. W., "Aids to Navigation and Meeting Traffic," U.S. Coast Guard Rept. CG-D-19-85, June 1985.

[14]Smith, M. W., Marino, K. L., and Multer, J., "Short Range Aids to Navigation Systems Design Manual for Restricted Waterways," U.S. Coast Guard Rept. CG-D-18-85, March 1985.

[15]Sennott, J. W., and Ahn, I. S., "Simulation of Optimal Marine Waypoint Steering with GPS, LORAN-C, and RACON Sensor Options," Fifth Annual Technical Meeting Institute of Navigation, Jan. 1988.

[16]Sennott, J. W., and Ahn, I. S., "Design of a State Estimator and Regulator for Marine Steering Incorporating Sensor Model Dynamics," *Proceedings of Twenty-First Annual Pittsburgh Conference on Modeling and Simulation,* May 1990.

[17]Gelb, A. (ed.), *Applied Optimal Estimation,* MIT Press, Cambridge, MA, 1974.

[18]Amerongen, J., et al., "Model Reference Adaptive Autopilots for Ships," *Automatica,* Vol. 11, 1975.

[19]Astrom, K. J., and Kallstrom, C. G., "Identification of Ship Steering Dynamics," *Automatica,* Vol. 12, 1976, p. 9.

[20]Fung, P., and Grimble M. J., "Dynamic Ship Positioning Using a Self-Tuning Kalman Filter," *IEEE Transactions on Automatic Control,* Vol. 28, No. 3, March 1983, pp. 339–349.

[21]Lewis, F. *Optimal Control,* Wiley, New York, 1986.

[22]Sennott, J. W., and Ahn, I. S., "Evaluation of Sensor Performance and Supervisory Control for Terminal Air Traffic," *Proceedings of Twenty-Second Annual Pittsburgh Simulation Conference,* May 1991, pp. 2135–2143.

Chapter 12

Applications of the GPS to Air Traffic Control

Ronald Braff*
MITRE Corporation, McLean, Virginia 22102
J. David Powell†
Stanford University, Stanford, California 94305
and
Joseph Dorfler‡
Federal Aviation Administration, Washington, DC 20591

I. Introduction

THIS chapter identifies and discusses applications of GPS to air traffic control (ATC). The first section provides a very brief overview of ATC for the purpose of providing the reader with sufficient operational context for understanding why and how the GPS will be integrated into the ATC system. The second section discusses basic operational requirements considerations with respect to GPS implementation into the ATC system. The third and fourth sections contain descriptions and discussions of the potential specific applications of GPS in providing navigation and surveillance services, respectively, to ATC.

II. Air Traffic Control System

Air traffic control is "a service provided by the appropriate authority to promote the safe, orderly, and expeditious flow of air traffic" (Ref. 1, p. 232). In the United States, the authority for ATC is the Federal Aviation Administration (FAA), which operates and regulates the National Airspace System (NAS). The NAS encompasses "the common network of airspace, airports, navigation aids, and air traffic control equipment across the United States" (Ref. 1, p. 536). Control of air traffic involves the following functions: 1) *procedures* and *regulations* by which the ATC system operates and the *organization of airspace* in the form of

Copyright © 1994 by the authors. Published by the American Institute of Aeronautics and Astronautics, Inc., with permission. Released to AIAA to publish in all forms.

*Principal Engineer, Center for Advanced Aviation System Development.

†Professor of Aeronautics and Astronautics and Mechanical Engineering.

‡Program Manager, Satellite Navigation Program.

routes for departures, en route airways, and arrivals; 2) *Human air traffic controllers* who are responsible for providing the ATC service; 3) *Automation systems* (e.g., computers and displays) providing information to the controllers on the status, location, and separation of aircraft in the system; 4) *Communications systems* providing air–ground and ATC interfacility voice and data communications; 5) *Surveillance systems* (e.g., radar) providing real-time positional information to ATC for tracking aircraft and hazardous weather; and 6) *Navigation systems* providing real-time positional information for aircraft navigation.

Today's continental airspace is mainly organized into airways defined by the VHF omnidirectional range (VOR) navigation aids. A VOR provides directional information to aircraft. The directional information is the angular parameter (referenced to magnetic north) of a polar coordinate system with the origin at the VOR antenna. The resulting radials of constant angle define the centerlines of VOR airways. To complete the polar coordinate positioning capability, distance-measuring equipment (DME) is collocated at most VOR facilities. An aircraft with only VOR capability (guidance on the selected radial) is constrained to fly VOR airways from one VOR to another, which extend into the terminal area. In most high-density terminal areas, the airways end at the approach feeder fixes, and air traffic controllers provide vectors (heading change commands) to line aircraft up on the final approach segment. Before the advent of small low-cost digital computers, the VOR was the best method for lateral guidance because the demodulation of its signals provides a direct read-out of the radial the aircraft is traversing.

The Instrument Landing System (ILS) provides both lateral guidance with respect to the extended runway centerline and vertical guidance for a fixed glide path of usually 3 deg. It also provides "marker beacons" at fixed points along the approach course to provide along-course information. The ILS is used in conjunction with lighting systems and other visual aids for final approaches during low-visibility conditions.

Area navigation (RNAV) permits aircraft to fly from waypoint to waypoint, where the waypoints can be defined at any location in two or three dimensions. A waypoint is a point in space that defines the beginning and end of a desired flight path. Radionavigation-based RNAV requires a computer to transform at least two signals into suitable guidance information. The essential benefit of RNAV is to allow users to choose, along a path defined by the waypoints, the best route they determine with respect to such performance criteria, as minimum time or fuel burn to destination. Examples of systems that provide RNAV capability are VOR/DME, DME/DME, Loran-C, Omega, the GPS, and the inertial navigation system (INS).

The present ATC system would have problems coping with the wide scale use of RNAV. The VOR airway system orders the flow of traffic, minimizing the number of route intersection points. With RNAV, aircraft would be flying user-preferred trajectories that would yield random routes, resulting in more complexity of the intersection of routes. This would be further compounded by the inevitable increase in traffic over the years. Hence, additional automation is needed to help controllers and traffic managers predict and resolve conflicts in such an environment; otherwise, the wide-scale introduction of RNAV would be impeded. The FAA will be implementing enhancements to the NAS Advanced

Automation System (AAS) that are designed to handle greater numbers of aircraft flying random routes. This set of enhancements is called automated en route ATC (AERA). An FAA benefit/cost study for AAS with AERA indicated that the accommodation of user-preferred trajectories accounted for nearly 70% of the benefits of AAS to the users.[2]

Figure 1[3] is a schematic diagram of the present NAS infrastructure of facilities. In addition, there are organizational entities providing standards and certification of user equipment and operation, and maintenance of the thousands of ATC facilities that provide the functions displayed in Fig. 1.

The foregoing description of the ATC system is the background to provide a context for the subsequent discussion of the application of GPS to ATC. An up-to-date detailed description of the ATC system is given in Ref. 1.

III. General Considerations

The role of the GPS in ATC involves providing highly accurate position, velocity, and time for the navigation and surveillance functions for all phases of flight and ground movement of aircraft. The GPS can be used for air navigation in three basic roles:

1) A *required navigation performance* (RNP) or *primary navigation* system is a navigation system that meets all requirements to use certain procedures or to fly in certain airspace without the need for any other navigation system onboard the aircraft, except, of course, compass and airspeed indicator. An RNP system may include one or more integrated navigation sensors in its definition (e.g., the GPS with an inertial reference system).

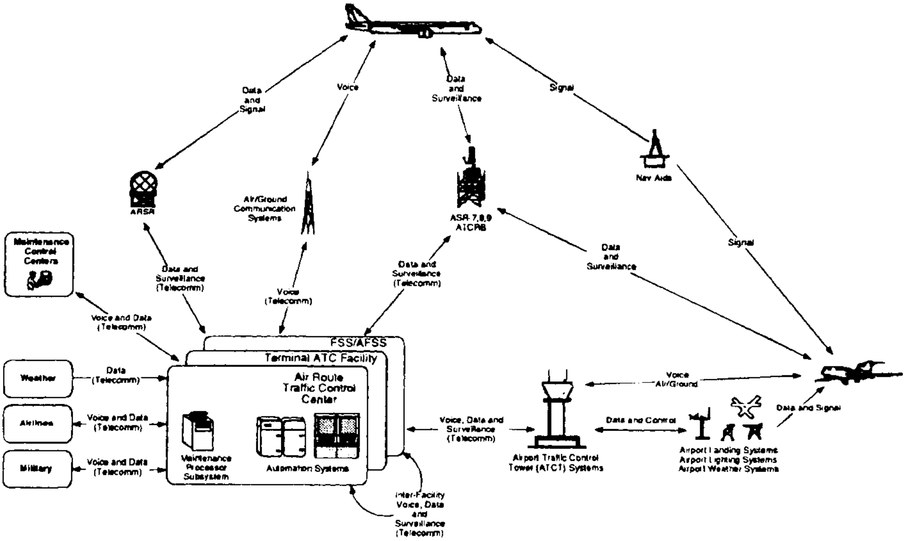


Fig. 1 1992 National Airspace System (NAS) infrastructure.

2) A *supplemental system* is one that can be used alone without comparison to another system; however, an RNP system that could be used in the event that the supplemental system is not available must be on board the aircraft.

3) A *multisensor navigation system* is one that can be used for navigation, but only after it has been compared for integrity with an RNP system in the aircraft.

When employed in surveillance, a navigation system can provide the sensor function that is used in conjunction with a data link to transmit positional reports to ATC. This type of surveillance is called "automatic dependent surveillance" (ADS).

Why is there such a great interest in use of the GPS in ATC? The basic benefits of the GPS are its higher accuracy and worldwide coverage of airspace. These benefits provide the user with the potential for minimum avionics that provides worldwide navigation capability to fly user-preferred routes, rather than airways, for all phases of flight, and the government an opportunity to reduce its vast infrastructure of thousands of ground-based transmitters dedicated to navigation and surveillance. Only a satellite navigation system such as the GPS provides all of these benefits. Specific user benefits include: capability to fly preferred routes, landing system capability to any runway, and reduced separation in nonradar airspace in conjunction with reliable air-to-ground datalinks for ADS-position reports. The benefits of satellite navigation and ADS are discussed in detail in Refs. 4 and 5.

Figure 2[3] illustrates a possible transition phase to GPS-based navigation and ADS. There is a mix of GPS-based and the ground-based systems where GPS is used in all phases of flight. Further transitions could involve complete GPS

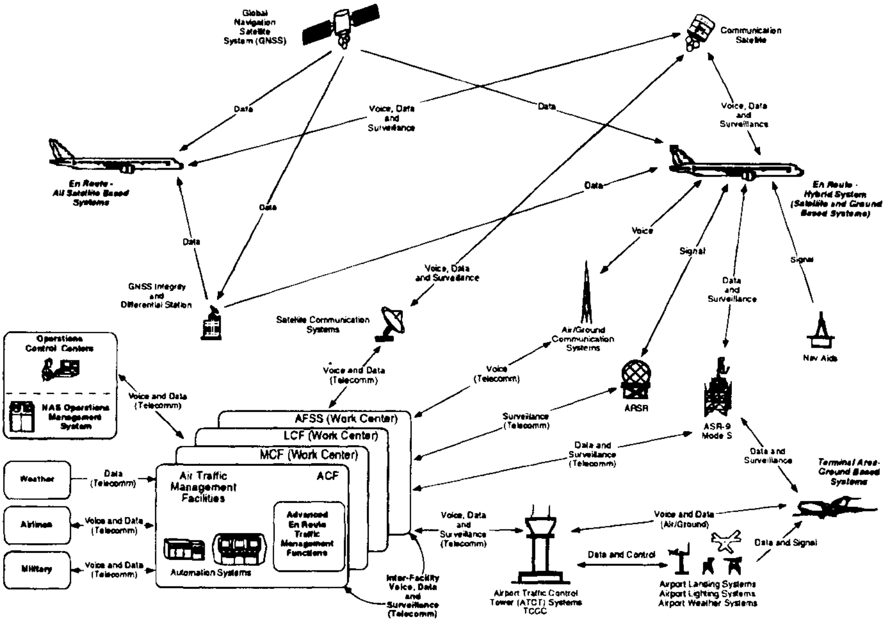

Global Navigation Satellite System (GNSS)
Communication Satellite
Data
Voice, Data and Surveillance
Data
Voice, Data and Surveillance
En Route - All Satellite Based Systems
Data
Data
En Route - Hybrid System (Satellite and Ground Based Systems)
Data
Voice
Signal
Signal
Voice, Data and Surveillance
Data and Surveillance
GNSS Integrity and Differential Station
Satellite Communication Systems
Air/Ground Communication Systems
Nav Aids
Operations Control Centers
NAS Operations Management System
Voice and Data (Telecomm)
Voice and Data (Telecomm)
Voice (Telecomm)
ARSR
Surveillance (Telecomm)
AFSS (Work Center)
LCF (Work Center)
MCF (Work Center)
Air Traffic Management Facilities
ACF
Advanced En Route Traffic Management Functions
Automation Systems
Weather
Data (Telecomm)
Airlines
Voice and Data (Telecomm)
Military
Voice and Data (Telecomm)
ASR-9 Mode S
Data and Surveillance (Telecomm)
Data and Surveillance
Terminal Area-Ground Based Systems
Voice, Data and Surveillance (Telecomm)
Voice and Data (Air/Ground)
Data and Signal
Data and Control
Inter-Facility Voice, Data and Surveillance (Telecomm)
Airport Traffic Control Tower (ATCT) Systems TCCC
Airport Landing Systems
Airport Lighting Systems
Airport Weather Systems

Fig. 2 2010 National Airspace System (NAS) infrastructure.

replacement of the ground-based navigation aids and much of the surveillance infrastructure. Such a stepped transition will be necessary because not all of the users can equip with the GPS in a short period of time, and ATC automation must evolve to accommodate and benefit from the increased capabilities provided by GPS-based navigation. As indicated in Fig. 3,[3] the approach taken here is to explore the role of satellites for navigation and ADS applications

A. Operational Requirements

The major operational considerations in the introduction of a new system concern its *accuracy, integrity, availability, continuity of service,* and the *procedures.*

Accuracy is the degree of conformance of estimated position with true position. It is usually expressed in statistical terms, such as the 95th percentile error. The most common accuracy metric for horizontal error is 2 drms, twice the rms radial (distance) error. If the navigation sensor errors are normally distributed then 2 drms is a circle about the true position containing approximately 95–98% of the position determinations, depending upon the eccentricity of the resulting bivariate error ellipse. Vertical accuracy is usually expressed at the 95 percentile or 2 standard deviation (2σ). Accuracy is usually the first parameter of consideration in the evaluation of a navigation system because it represents a physical limitation of a system, and it is straightforward to estimate by analysis or measurement.

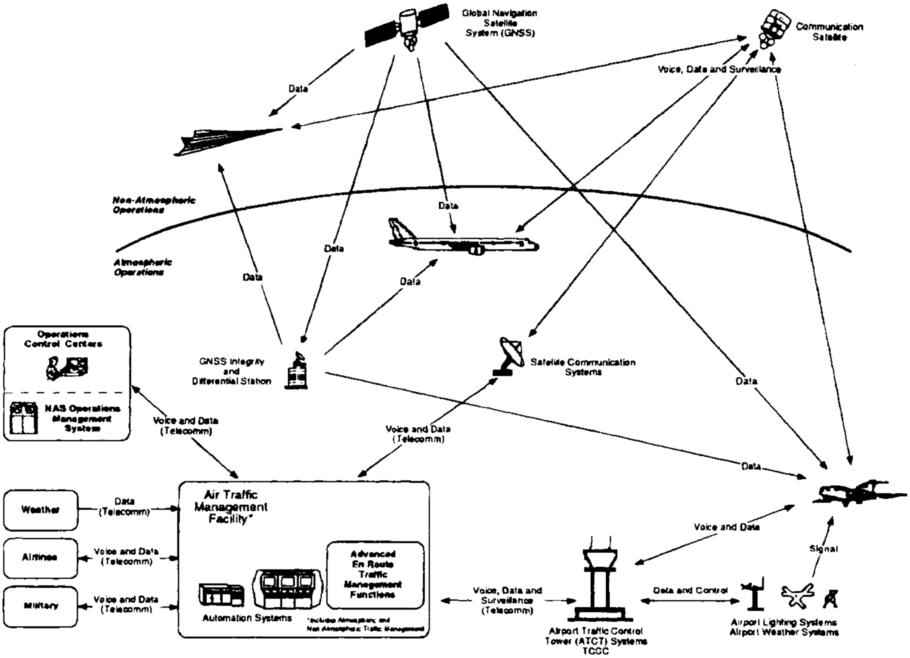

Global Navigation Satellite System (GNSS)
Communication Satellite
Voice, Data and Surveillance
Data
Non-Atmospheric Operations
Atmospheric Operations
Data
Data
Data
Data
Operations Control Centers
NAS Operations Management System
GNSS Integrity and Differential Station
Satellite Communication Systems
Data
Voice and Data (Telecomm)
Voice and Data (Telecomm)
Data
Weather
Data (Telecomm)
Air Traffic Management Facility*
Airlines
Voice and Data (Telecomm)
Military
Voice and Data (Telecomm)
Advanced En Route Traffic Management Functions
Automation Systems
*Includes Atmospheric and Non Atmospheric Traffic Management
Voice, Data and Surveillance (Telecomm)
Airport Traffic Control Tower (ATCT) Systems TCCC
Data and Control
Voice and Data
Signal
Airport Lighting Systems
Airport Weather Systems

Fig. 3 20?? Air/space traffic management system infrastructure.

"*Integrity* is the ability of a system to provide timely warnings to users when the system should not be used for navigation."[6] The integrity function of a navigation system involves monitoring of the system's errors, and if specified protection levels are estimated to be exceeded, a warning is given to the pilot that the system cannot be used for navigation, or the system shuts itself off.

The integrity requirements and solutions with respect to a combined use of the GPS as the sensor for both navigation and ADS has not yet been addressed. In the present radar-covered airspace where the navigation and surveillance functions are completely independent, there is a very high level of integrity in the sense that although the ground-based radionavigation systems have an excellent integrity function, any failure of that function would be caught by the independent radar surveillance system. Furthermore, if an aircraft loses its ,radionavigation capability when in radar airspace, then its dead-reckoning navigation can be monitored by radar, and speed and heading corrections ("radar vectors") communicated by the air traffic controller to the pilot via voice radio transmissions. The issue here is whether more demanding integrity requirements should be put on the GPS when it is both a sensor for navigation and ADS. The introduction of more stringent integrity requirements could entail an increased alarm rate, thereby decreasing availability. Any time there would be an integrity alarm, there could be an absence of both navigation and surveillance to the affected users.

"The *availability* of a system is the percentage of the time that the services of the system are usable."[6] The FAA requires an availability of 99.999%[7] for RNP enroute and terminal navigation, and for surveillance. This only applies to the services provided by the FAA and does not include the airborne equipment. When serving high-density airspace, this high availability is attained through redundant coverage of navigation and surveillance ground facilities, and redundant subsystems within each facility. The 99.999% requirement has been used as a guideline in estimating augmentations to increase the GPS availability.

The separation of routes in nonradar covered airspace is much greater than in radar airspace (e.g., 60 n.mi. vs 8 n.mi.). Thus, the first combined application of GPS navigation and GPS-based ADS with adequate datalinks will be in oceanic airspace where there is a great desire to decrease the large required separation between aircraft tracks. Using qualitative arguments, the formerly Radio Technical Commission for Aeronautics (RTCA) Task Force report (Ref. 4, p.17) was quite optimistic about the feasibility of adequate integrity and availability of GPS applications for all airspace. GPS availability requirements should not be as stringent in nonradar airspace as in radar airspace because there is a wide margin for decreasing aircraft separation standards, but still keeping them larger than in radar airspace.

Continuity of service is the ability of a navigation system to provide required service over a specified period of time without interruption. Continuity is particularly important in the approach and landing phase of flight. "The level of continuity is expressed in terms of the probability of not losing the radiated guidance signals."[8]

Procedures are based on criteria that have to do with where and how the system can be certified for operation. This is determined by the regulatory authority (FAA) based on the system's capabilities. For example, at the time of writing, GPS has been certified as a supplemental system for en route and nonprecision approach navigation.[9] A nonprecision approach procedure is that wherein a radionavigation system provides only lateral guidance for the approach; whereas in a precision approach procedure, the radionavigation system provides both lateral and vertical guidance.

When a system is implemented with new and better capabilities, the users would like to use it to its full capabilities to derive the most benefits. However, mainly for safety reasons, it takes regulatory authorities time to approve use of a system to its fullest capabilities. For example, the GPS standard positioning service (SPS) accuracy (100 m, horizontal) is such that it could be used to provide a level of accuracy for nonprecision approaches that is at least equivalent to, or, in most cases, better than that provided by today's standard, the VOR.[6] However, to take full advantage of this capability, new criteria must be developed so that procedures can account for smaller ground obstacle clearance criteria for a GPS-based system. Therefore, as a transition strategy, the FAA is allowing suitably equipped aircraft[9] to use the GPS to fly all present nonprecision approaches, except those where lateral guidance is provided by the localizer subsystem of the ILS.

B. Government Activities

The following description of government activities concerning the GPS's application to ATC is presented to indicate the seriousness of the commitment to lay the ground work for the GPS to assume a major role in the NAS and internationally.

The FAA Satellite Navigation Program Plan,[10] updated annually, provides the scope, objectives, schedules, and other requisite planning information for implementing satellite navigation in the NAS. The FAA program covers all required activities for implementation of satellite navigation in all phases of flight, including precision approaches.

The FAA Satellite Operational Implementation Team (SOIT) was formed on 19 August 1991 to facilitate the introduction of satellite navigation and communications into the NAS. The team consists of FAA experts in aviation and flight standards, avionics certification, instrument flight procedures, and other operational areas. At the time of writing, the team has developed and approved the process, procedures, and standards for operational use of the GPS for all phases of flight down to nonprecision approaches: on 25 February 1991 (and revised 20 July 1992) for using the GPS as an additional sensor input to an approved multisensor navigation system; and on 10 December 1992 a Technical Standard Order (TSO)[9] was written describing the required capabilities of GPS receivers to be used as a supplemental system of navigation. Furthermore, on 23 April 1992, the FAA published the following notices developed by the SOIT: Notice 8110.47, "Airworthiness Approval of GPS Navigation Equipment for Use as a VFR and IFR Supplemental Navigation System," and Notice 8110.48, "Airworthiness Approval of Navigation or Flight Management Systems Integrating Multiple Navigation Sensors." On 9 June 1993 the FAA authorized supplemental navigation approval for use of GPS equipment (TSO-C129) to conduct oceanic, domestic en route, and terminal instrument flight rules (IFR) operations, as well as nonprecision approaches with certain limitations.

RTCA activities relating to the GPS have been underway since the establishment of RTCA Special Committee 159 (SC-159), at the request of FAA, on 20 September 1985. The RTCA provides an organizational framework for interested parties representing airspace users, avionics manufacturers, and government organizations who volunteer to develop, by consensus, minimum operational performance standards (MOPS) for avionic systems, and more broad-based consensus on defining communication, navigation, and surveillance (CNS) systems for aviation use and determining their benefits.[4] The SC-159 documents are recommended guidelines used by the FAA and other parties in developing TSOs and technical programs. The SC-159 MOPS for supplemental GPS[11] navigation equipment was used as a major technical input for the TSO-C129 on the GPS for supplemental navigation. At the request of the FAA, the RTCA also formed a task force to develop a consensus strategy with recommendations regarding early implementation of an operational Global Navigation Satellite System (GNSS) capability in the United States. The task force was composed of high-level representation from commercial, business, and general aviation users, industry, the U.S. Department of Defense, and the FAA. They reached a solid consensus that the user community wants, needs, and is ready to implement GNSS-based operations, and that the benefits will apply to virtually all aspects of aviation operations.[4] GNSS is a concept for "a worldwide position and time determination system. GNSS includes one or more satellite constellations, end user receiver equipment, and a system integrity monitoring function. GNSS will be augmented as necessary to support the RNP concept for a wide range of specific operations" (Ref. 4, p. 9).

The *International Aviation Civil Aviation Organization* (ICAO) Special Committee on Future Air Navigation Systems (FANS) has defined a future ATC system where satellites play the major role in providing the communications, navigation, and surveillance (ADS based on GNSS) infrastructure.[12] The ICAO is an organization of the United Nations responsible for promulgating standards and recommended practices (SARPS) that have the status of international treaties. The purpose of SARPS is to ensure the international interoperability of CNS systems. Because the GPS is an integral part of the envisioned GNSS, it receives attention at ICAO. In 1992, the U.S. government stated to ICAO its intention to provide GPS signals for the foreseeable future with no direct user charges (Ref. 4, p. 57). It was also stated that ICAO be afforded at least 6 years' advance notice prior to the termination of GPS signals.

The *Federal Radionavigation Plan* (FRP) "delineates policies and plans for federally provided radionavigation services."[6] It has a biannual update. The FRP includes the Federal government's policy on GPS and information on the retention of other radionavigation systems that may be impacted by GPS. It is a collaborative effort by DOD and DOT; therefore, it includes information relevant to aviation and marine- and land-based users, both in the civilian and military communities.

IV. Air Navigation Applications

Applications of the GPS to air navigation are best partitioned into 1) en route, terminal, and nonprecision approach phases of flight; and 2) precision approach phase of flight. This is a natural partition for historical and practical reasons.

The 100-m (2-drms) accuracy specification for the GPS standard positioning service is based on civil aviation's need for a level of accuracy that is as good or better than present approved navigation systems for all phases of flight down to and including nonprecision approaches. Therefore, in the early 1980s, this SPS accuracy was recommended by the FAA and accepted by the Department of Defense. At that time, receiver and differential GPS (DGPS) technology and projected system performance precluded GPS application to precision approaches where very accurate vertical guidance is required.[13] However, within the last few years flight tests of DGPS, with state-of-the-art receiver developments, have indicated the accuracy feasibility of GPS for precision approaches. From the users' point of view, the GPS may be looked upon as a "seamless" potential replacement of inertial navigation and Omega for oceanic en route navigation, VOR/DME for domestic navigation through nonprecision approaches, and ILS at least up to Category I (CAT I) approaches. At the time of writing, application of the GPS to the more stringent CAT II and CAT III categories of approach are under intense investigation.[14]

The following discussion stresses the navigation operational considerations for GPS. This discussion should provide the reader with some insight as to what is involved in implementing the GPS as an approved air navigation system.

A. En Route, Terminal, and Nonprecision Approach Operational Considerations and Augmentations

1. Accuracy

The accuracy of an air navigation system usually considers three basic error sources; namely, sensor error, course-centering error, and flight technical error (FTE). The sum of these three errors is called total system error (or system use error). For clarity, these errors are defined here by the way they are measured.

1) *Sensor error* is the difference between the navigation receiver position determination and a truth source of position (such as a surveyed point, theodolite angle, or a laser tracker position). Sensor error consists of the sum of the nominal GPS error components (e.g., signal, atmospheric delays, receiver noise, multipath, and coordinate conversion) (Ref. 11, p. 15). Sensor error is usually expressed as twice the rms error (or the error not exceeded 95% of the time).

2) The *FTE* is the measure of how well a human pilot or an autopilot can follow the guidance commands derived from the navigation position determination. It is measured as the negative of the guidance command. It is usually expressed as twice the rms error or a 95% error. It considers pilot performance in specified wind environments. The FTE must be such that it will not cause an unacceptable total system error when combined with sensor accuracy.

3) *Course-centering error* is the measure of how accurately the navigation sensor position is transformed into guidance commands, where the guidance commands are relative to the desired flight path. It is the difference between the displayed cross-track guidance command and the computed guidance command.

4) *Total system error* has been traditionally calculated as the root-sum-square of the 2-rms sensor error, course-centering error and FTE. More recently, for approach flight tests, the FAA has also been estimating total system error as the difference between a laser tracker truth position and the assigned flight path.

There should be, and it is generally observed, that there is no significant difference in the results obtained by the two methods.

The cross-track accuracy requirements for each of the three phases of flight through nonprecision approach have been agreed upon (Ref. 11, p. 31). The 95% sensor error requirement for GPS RNAV is 0.124 n.mi. (230 m) for all phases of flight through nonprecision approach. The course-centering error requirement for the en route and terminal phases of flight are 0.2 n.mi. (370 m), and 0.1 n.mi. (185 m) for nonpecision approach. No FTE requirements are included in Ref. 11 because it is stated that FTE is beyond the control of the equipment manufacturer or installer. Estimates of manual FTE are 1.0, 1.0, and 0.5 n.mi. for domestic en route, terminal, and nonprecision approach, respectively. For coupled automatic flight control the FTE estimates are 0.25, 0.25, and 0.125 n.mi. for domestic en route, terminal and nonprecision approach, respectively.[11] Root-sum-square (rss) combining of the aforementioned error budget components yields estimates of 95% total system error contained in Table 1.

The aforementioned accuracy values are expected to hold for both supplemental and RNP GPS navigation.

2. *Integrity*

Integrity of the GPS as a navigation system in the NAS has been addressed and recognized as an issue for many years (Ref. 15, pp. 1214–1223 and Ref. 11, Appendix B). The problem is that GPS as implemented today does not have the capability to notify users of a signal malfunction in a timely manner. For some types of signal malfunctions, it can take on the order of an hour for notification; whereas, the integrity monitoring response times required for flight operations are on the order of seconds. Integrity solutions have been found and are described elsewhere in this volume.

The provision of integrity for supplemental satellite navigation requires only the detection of a navigation sensor malfunction; whereas, the integrity of an RNP system requires both detection and correction of the malfunction. The latter requirement is needed for RNP because a faulty satellite needs to be identified so it can be removed from the position determination solution, or a subset of satellites that do not cause an alarm can be found. There are two basic ways for providing integrity: ground-based monitoring (the GPS Integrity Channel or GIC), and airborne monitoring by using redundant measurements (Receiver Autonomous Integrity Monitor or RAIM).

The GIC will consist of a network of ground-based GPS signal-monitoring stations located at known reference points that cover a wide geographical area

Table 1 Total system cross-track error estimates for GPS

Phase of flight	Steering	95% Total system error, n.mi.
Domestic en route	Manual	1.0
	Coupled	0.3
Nonprecision approach	Manual	0.52
	Coupled	0.2

over which signal integrity is guaranteed by a navigation provider, such as the FAA. These monitors will be connected to a central control station where the integrity decisions will be made and messages composed. The integrity messages will be broadcast through geostationary satellite relays. Recently the FAA has called its implementation of the GIC the GPS Integrity Broadcast (GIB).

Receiver autonomous integrity monitoring (RAIM) (see Chapter 5, this volume) is essentially various algorithmic techniques for integrity monitoring that use redundant pseudorange measurements (i.e., $n - 4$ satellites when $n > 4$ satellites are visible) or aiding from another sensor (e.g., barometric altimeter). Receiver autonomous integrity monitoring without aiding requires at least five satellites in view with good geometry to permit detection of a violation of GPS position error tolerance. At least six satellites with good geometry are required for the identification of a faulty satellite or the determination of a useful subset of satellites.

With the planned GPS constellation and no augmentations to GPS, the availability of RAIM identification for RNP-nonprecision approach is clearly not sufficient for operations. Because identification is required for RNP, the FAA is considering the implementation of a GIC to provide the integrity monitoring function.[16] Not only is GIC expected to provide enhanced integrity monitoring, it will also increase availability and continuity of service because navigation can be conducted with only four satellites in view. The requirements for integrity monitoring are usually stated in terms of four parameters.[11]

a. Protection level. The positional error magnitude that cannot be exceeded, and for which the integrity monitoring system provides protection by warning the pilot.

b. Alarm rate. An alarm must be annunciated when position errors exceed the protection level. The alarm rate should not be so excessive that it becomes a nuisance during operations.

c. Time to alarm. The maximum allowable time from the onset of a failure to the annunciation of the failure.

d. Missed detection probability. The probability that a failure occurred and was not detected and displayed to the pilot.

Table 2, derived from Ref. 11, contains the RTCA SC-159 recommended requirements for RAIM integrity monitoring for supplemental navigation in the en route, terminal, and nonprecision approach phases of flight.

The RAIM performance requirements for RNP have not been specified at the time of writing, but they may contain a much lower maximum allowable alarm rate to ensure the high performance required for an RNP system. When RAIM identification is possible, no alarm need be annunciated if a fault is identified, and sufficient navigation capability is provided by the other satellites in view. Integrity monitoring requirements must be specified for the GIC, a point that is discussed later with respect to the FAA's proposed GIC implementation.

3. Availability

The availability of GPS for navigation has been addressed for many years.[16] Assessments of GPS availability for air navigation indicate that use of the planned

Table 2 GPS RAIM performance requirements for supplemental navigation as a function of phase of flight

Phase of flight	Protection level, n.mi.	Maximum allowable alarm rate, h^{-1}	Time to alarm, s	Minimum detection probability
En route	2.0	0.002	30	0.999
Terminal	1.0	0.002	10	0.999
Nonprecision approach	0.3	0.002	10	0.999

constellation, without any augmentations from other navigation sensors, may fall short of providing sufficient availability in the context of GPS as an RNP system, especially for the approach phase of flight.[17,18] These referenced studies used average availability over continental airspace as the measure of GPS availability. However, because the geometry of GPS satellites is varying in both time and space, these variations must also be considered. Figure 4[19] illustrates the diurnal variation of unavailability (1—availability) of a 24-satellite GPS constellation augmented with Inmarsat-3 geostationary satellites, where both the geometric constraints on satellite–user geometry and satellite failures are considered. Note the significant order-of-magnitude changes in unavailability during the day. Clearly, this variation indicates that average availability is not a complete measure. The times where the spikes indicate poor availability could be eliminated through further augmentations. Figure 5[19] illustrates the spatial variability of GPS (augmented by Inmarsat-3 satellites) unavailability using average unavailability as the measure. Again, significant variation can be seen.

The availability of a satellite-based system is of more concern than that of a ground-based system because a loss of signal coverage in the former could involve a very wide area; whereas, a ground-based system outage would involve a facility that covers a much smaller area. For example, the impact of an outage of ILS could be minimized by diverting aircraft to another ILS runway on the airport, if feasible, or to a nearby alternate airport. However, with a satellite-based system, the alternate airport could also have lost the signal coverage. On the other hand, the ILS outage persists until repaired, which could take several hours from intitial equipment shutdown; whereas, a GPS outage caused by lack of satellite coverage or sufficient postion fix geometry could be alleviated within 20 minutes or so when another satellite(s) comes into view. Also, the predictability of satellite availability at the destination (made at departure time) would ease the impact of periodic satellite coverage holes caused by failed satellites.

A predicted outage does not compromise safety, but it would have an impact on the efficiency of the ATC system if GPS-RNP is to replace existing navigation aids. These are the reasons why augmentations to the GPS are under consideration, such as hybrid combination with other systems (e.g., GLONASS and Loran-C), and additional satellites with GPS-like signals.[10,20-22] These augmentations of the GPS to increase availability are discussed below.

The requirements for availability depend upon whether the GPS is to be used as a supplemental or an RNP system. Reference 9 contains no quantitative signal

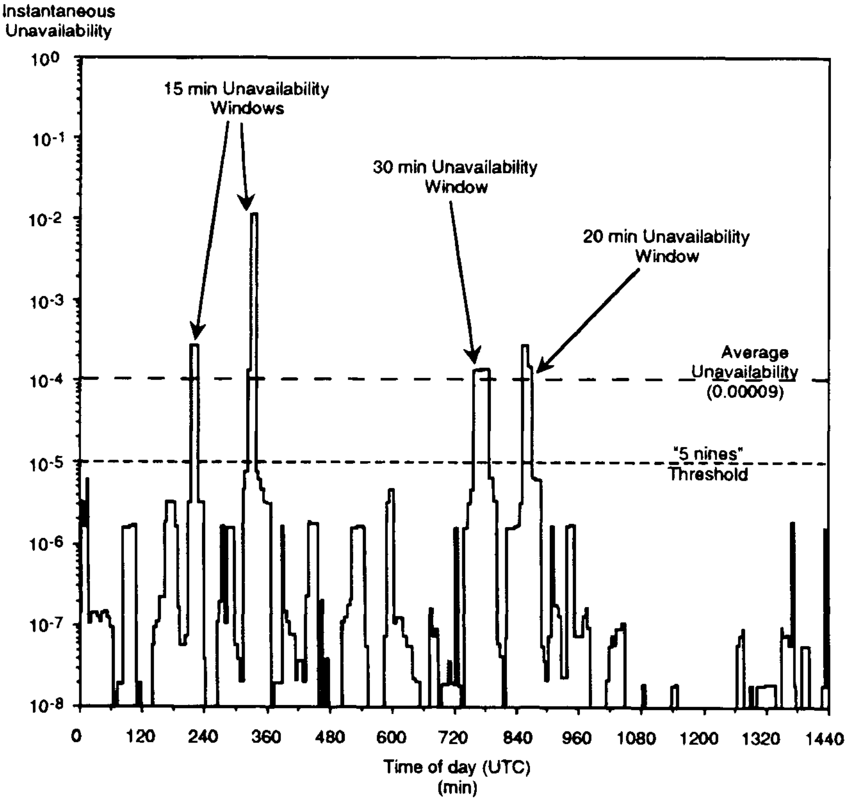


Fig. 4 Temporal unavailability: GPS + Inmarsat overlay at Dallas, TX (VDOP > 4.5).

availability requirements for supplemental GPS. However, availability of the GPS for supplemental use must to have a reasonable value (e.g., 95%); otherwise, it could not be relied upon to provide adequate area navigation service, and could be disruptive to air traffic control when aircraft have to transition from area navigation to VOR radial navigation during a GPS signal outage. Stated availability requirements for supplemental use tend to be concerned with predicting availability of RAIM (sufficient number of satellites and geometry), and providing enhancements to increase it. For instance, there is a requirement for the GPS navigation set to provide the pilot with information to determine whether RAIM will be available at the planned destination.[9] To enhance availability, there is also a requirement for barometric altimeter aiding of the RAIM function.

At this time, there seems to be no stated availability requirement for GPS as an RNP system. However, it seems that such a requirement may have to consider the 99.999% requirement stated in Sec. II. Augmentations to GPS that have the potential to achieve this level of availability are discussed below.

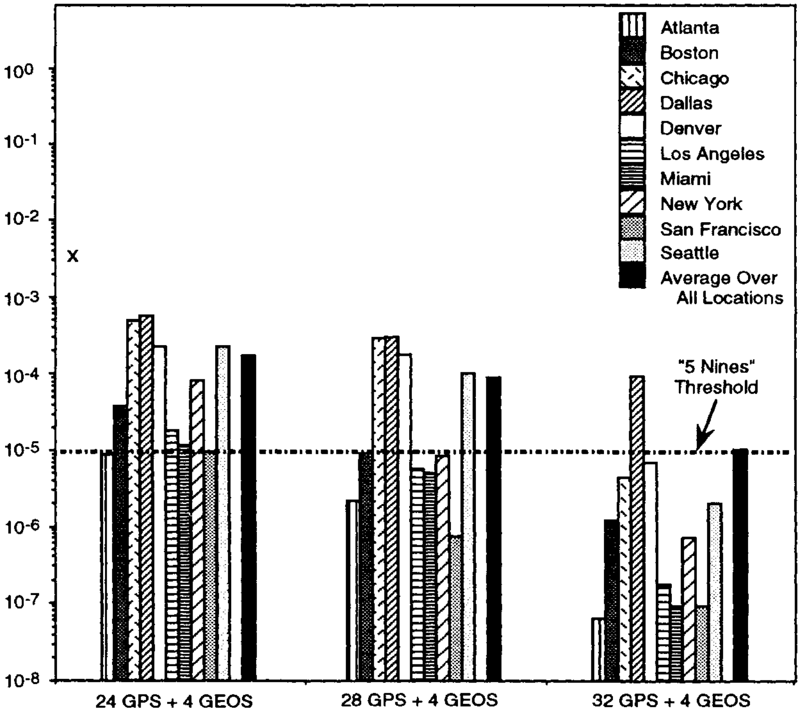


Fig. 5 Spatial unavailability: Various Constellation Sizes of GPS + Inmarsat Overlay (VDOP > 4.5).

4. Continuity

For GPS supplemental continuity, the traditional requirements tend to be somewhat relaxed. For instance, during the final approach segment of a nonprecision approach, with the GPS navigation function in operation, the warning flag is not displayed until the RAIM detection function is lost for more than 5 min.[9] Also, if RAIM is augmented by altimeter or receiver clock coasting, the warning flag may be delayed for a period of time consistent with the worst case drifts of these aiding sensors.

There are no existing quantitative or qualitative requirements for RNP continuity for nonprecision approach; however, a sometimes stated rule of thumb is no more than 1 out of 1000 approaches should be broken off because of loss of navigational signals.

5. Augmentations

The following are discussions of some augmentations for increasing GPS availability with respect to an adequate integrity function. The first three augmentations, barometric altimeter, Loran-C, and GLONASS, increase the availability of RAIM. The GIC with geostationary satellite overlay, the last discussed augmentation, provides an external ground monitoring system and communication of GPS signal integrity to users. Relying on sources available in the navigation literature, an estimate of the availability of the integrity function attributable to each augmentation is provided. These availabilities are not compared in a quantitative sense because they were estimated by different parties using somewhat different assumptions and methods. However, as shown, the availability estimates seem to indicate that GIC with geostationary satellite overlay and GPS/GLONASS would provide sufficient availability of the integrity function to provide RNP nonprecision approach.

a. Barometric altimeter. Barometric altimeter augmentation is a relatively inexpensive method of augmentation. In the FAA TSO-C129 for supplemental navigation,[9] it is required to be used in conjunction with RAIM for the purpose of increasing RAIM availability. A digital altimeter read-out is the input to the navigation solution, along with the pseudoranges. The altimeter may be looked upon as a ranging source from the center of the Earth when the altitude readout is converted into a range from the center of the Earth using the mean sea level relationship to the ellipsoid. The barometric altitude read-out can be calibrated by either of two methods.

One method is to calibrate the altimeter with the local pressure correction that is transmitted from an airport. This ties the altitude to the surface of the airport. The other method is to calibrate the altimeter with GPS-derived altitude while geometry is adequate to ensure integrity of the GPS vertical position. Present approved practice is to use the altimeter input only as a source for RAIM when RAIM is not available; e.g., only four satellites are in view, or the RAIM geometry is inadequate.[23] Reference 23 contains the equations for analyzing altimeter inputs.

Table 3 presents some results for en route availability of RAIM for horizontal position error protection levels of 1 and 2 n.mi.[23] These results assume selective availability (SA) is in place (pseudorange error standard deviation = 33 m); 21 operating satellites in the the GPS constellation; and satellites are used if they have an elevation angle above 7.5 deg (or 7.5 deg mask angle). It is seen that even with a typical set of 21 satellites up, the RAIM availability for en route

Table 3 21-satellite constellation RAIM detection availabilities over some air routes, %

Augmentation	Route	Protection level = 1 n.mi.	Protection level = 2 n.mi.
None	New York–Los Angeles	88.5	91.3
None	San Francisco–Japan	91.0	93.7
None	Dallas–Paris	92.9	95.0
Altimeter via GPS	New York–Los Angeles	93.9	94.9
Altimeter via GPS	San Francisco–Japan	94.3	95.7
Altimeter via GPS	Dallas–Paris	96.0	98.1

supplemental navigation is generally adequate for a protection level of 2 n.mi. For that protection level, the availability is generally in the neighborhood of 95%.

Table 4 presents the results for nonprecision approach extracted from Ref. 23. These results assume SA is in place (pseudorange error standard deviation = 33 m); barometric altitude correction standard deviation = 49 m; a typical set of 21 operating satellites in the GPS constellation; and 7.5 deg mask angle. It should be recognized that the absolute availability values are strongly dependent on the foregoing parameter assumptions, particularly the mask angle. However, the results provide an insight into the relative benefits of the two methods of calibrating a barometric altimeter. Referring to the fourth column of Table 4, it is seen that for nonprecision approach GPS calibration increases availability for navigation, but has negligible effect in increasing RAIM detection availability. However, local calibration does provide a significant increase in RAIM availability.

b. Loran-C augmentation. Loran-C is a low-freqency navigation aid that covers the conterminous U.S and is certified as a supplemental air navigation system. It is normally operated in the hyperbolic mode, where each line-of-position is derived as the difference between the time-of-arrival of two signals. A detailed description of Loran-C is given in Ref. 24.

It has been recognized that combining Loran-C with the GPS can provide an availability for navigation that is significantly greater than that of either system

Table 4 21-satellite constellation RAIM availabilities for nonprecision approach at some major airports, % (protection limit 0.3 n.mi.)

Augmentation	Airport	Navigation only	RAIM detection of position error	RAIM identification of a malfunctioning satellite
No augmentations	SFO[a]	100	72.2	28.8
	DFW[b]	100	68.1	25.7
	ORD[c]	100	70.1	31.3
	JFK[d]	97.0	68.8	30.2
	ATL[e]	96.5	73.3	32.3
GPS calibration of altimeter	SFO	100	72.6	42.0
	DFW	100	69.1	36.8
	ORD	100	71.5	43.8
	JFK	99.0	69.1	46.5
	ATL	99.7	74.0	46.5
Local calibration of altimeter	SFO	100	86.5	68.8
	DFW	100	83.7	66.0
	ORD	100	89.2	67.0
	JFK	100	85.1	68.4
	ATL	100	88.2	68.4

[a] San Francisco.
[b] Dallas-Forth Worth.
[c] Chicago.
[d] Kennedy-New York.
[e] Atlanta.

alone, but more importantly, it will increase the availability of RAIM. In Ref. 25, descriptions are given of how the GPS could be augmented with Loran-C signals. The essence of this concept is to treat a Loran-C signal as a pseudorange, and combine the Loran-C pseudoranges with the GPS pseudoranges. However, the minimum number of pseudoranges required depends upon whether Loran-C system time is accurately calibrated with the GPS system time. If time is not accurately calibrated, then one extra pseudorange measurement is required in the position determination solution. It is estimated that such a combined system would have access to at least 9 and an average of 11 pseudorange measurements.[15]

Reference 25 contains curves showing the unavailability of GPS/Loran-C hybrid RAIM (21-satellite constellation) as a function of a position determination geometry parameter. For the geometries of the New England/New York area, the unavailabilities are essentially constant, independent of that parameter. The results indicate the availabilies for combined GPS/Loran-C fault detection and isolation increase to the orders of 99.99 and 99.9%, respectively. This is a very significant improvement over GPS-RAIM (see Table 5). Thus, this hybrid combination shows great promise for increasing RAIM availability.

c. GLONASS augmentation. GLONASS is a Russian navigation system that is planned to be a 24-satellite constellation. A description of GLONASS and its performance are given in Ref. 26, where it is stated that there seems to be no problem in bringing the spatial and time coordinates of the two systems into coincidence.

With a combined GPS/GLONASS, each containing 21 operating satellites, 99% of the users would see 10 or more satellites and almost all would see at least 8 satellites.[27] Thus, augmentation with GLONASS would provide a great increase in RAIM availability. Evaluation of a RAIM algorithm, assuming 21 operating satellites in each constellation, indicated that for a critical-system availability of 99.999%, a horizontal protection level of approximately 350 m (0.2 n.mi.) could be achieved. Thus, a combined GPS/GLONASS satellite navigation system shows great promise in providing an RNP nonprecision approach capability.

d. GPS integrity channel. Reference to column 3 in Table 4 indicates that for nonprecision approach, there is near 100% availability for navigation if RAIM were not required. With the implementation of the GIC, these levels of availability will be achievable and perhaps surpassed, as explained below.

When the the GIC was first envisioned, the health of each GPS satellite signal was determined by the ground integrity monitoring segment, and the GIC broadcast indicated the healthy satellites.[28] If a satellite were to be indicated unhealthy, it was not used in the position solution. The broadcast would be via planned mobile communication satellites. Later, however, the RTCA Special Committee SC-159[29] developed a concept where the GIC would broadcast the satellite errors that were estimated by the ground integrity monitoring segment.

The integrity message, consisting of satellite error estimates, would be broadcast as part of the navigation message of the GPS-like signals transmitted from geostationary satellites, and the signals also would provide additional sources of pseudoranges.[21] The quantization of the broadcasted errors would be coarse (e.g., preset 25-m quantization levels). The airborne receiver would then use these errors as an input to a real-time position error estimation process that modeled the satellite geometry being used. If the estimated position error remained within the horizontal position error protection level, integrity of the signals used in the GPS position determination would be assumed. If not within the protection level, an attempt would be made to find a combination of satellites yielding a solution that was within the level; otherwise, an alarm flag would be raised.

The advantage of the RTCA approach is that it would produce a much lower alarm rate because it tailored the integrity decision to the user's satellite geometry and integrity protection based on phase of flight. During this period, it was decided not to use the GIC-estimated errors as differential corrections because of the traditional reluctance of mixing postion determination with integrity monitoring.

Presently, the bold step of using the GIC error estimates as corrections is under serious consideration. This step may be justified by realizing that if effective independent ground monitoring of the corrections can be achieved, then the corrections would be guaranteed to provide a truncation of any large signal-in-space errors. In this sense, the corrections can be considered as integrity-monitored "digital navigation signals" rather than as integrity messages so that the tradition of separating position determination from integrity is not violated. The effective integrity monitoring of the corrections can be achieved by independent ground monitors that compare their estimates of the pseudorange error components (e.g., satellite clock, orbital data, and ionospheric errors) with those received from the GIC broadcast.

The bounding of signal-in-space errors bounds the resulting position determination errors. It was shown in Ref. 30 that there is a factor, HMAX, so that $\delta r < \mathrm{HMAX}_n \cdot E$, $|e_i| < E$, $i = 1,2, \ldots, n$, where δr is the horizontal radial error; E is a known bound on the range errors; $|e_i|$ are the magnitudes of the pseudorange measurement errors; and n is the number of satellite pseudoranges used in the position solution. Assuming the errors attributable to the quantization of the coarse corrections are the major source of pseudorange errors attributable to signal-in-space errors (within ± 12.5 m for a 25-m quantization), the signal-in-space measurement errors can be assumed to be bounded by the half-quantization interval. It then follows that E represents the half-quantization interval.

Also, for a four-satellite position determination solution, $\mathrm{HMAX}_4 < 2\ \mathrm{HDOP}_4$,[30] where HDOP is horizontal dilution of precision (a measure of the amplification of position error attributable to satellite geometry), $\delta r < 2\ \mathrm{HDOP}_4 \cdot E$. The utility of this bound is that most accuracy analyses of the GPS are given in terms of HDOP. For example, if $\mathrm{HDOP}_4 < 4$ (a large upper bound), and for 25-m quantization, the resulting position error upper bound (attributable to the signal-in-space) is 100 m. Reference 31 contains results of an analysis indicating that 0.99999 availability would be obtained for $\mathrm{HDOP} < 4$ if GPS is used in conjunction with geostationary satellite signal sources.

B. Precision Approach Operational Considerations and Augmentations

Differential GPS (DGPS) will make it possible to conduct precision approaches at any runway within the coverage area. As pointed out in Ref. 10, important safety benefits will result because both horizontal and vertical centerline guidance will be provided throughout the approach, requiring no alignment or transition. Moreover, if aircraft can be directed to all available runways during instrument meteorological conditions (IMC), significant improvements in capacity and terminal airspace capacity will be possible. Differential GPS may also be used for airport surface position determination and situational awareness.

Precision approach is the most demanding application of GPS to air navigation. First, the sensor accuracy requirements are of the order of 20 times more stringent as for nonprecision approach, and it involves the need for very accurate vertical guidance. It is well known that GPS vertical accuracy is usually significantly less than horizontal accuracy (e.g., the average ratio of VDOP/HDOP = 1.4[6]). Second, the position update rate requirement is much higher, especially for autopilot coupled approaches. Third, the integrity and continuity requirements are the most stringent. Thus, for precision approach some type of differential augmentation is needed to achieve very accurate position and velocity determinations.

1. Accuracy

The lateral and vertical sensor accuracies for precision approach traditionally have been based on the three categories of approach: Category I, II, or III (or CAT I, etc.) The operational definitions of these categories are based on visibility or runway visual range (RVR) and decision height (DH). There is equipment at airports (called RVR system) that measures the visibility along the runway, and this measurement is reported to pilots approaching the airport. The CAT III requirements also include very stringent equipment redundancies, lateral guidance in rollout for CAT IIIb and CAT IIIc, and other requirements.[32] The DH is "a specified height at which a missed approach must be initiated if the required visual reference to continue the approach to land has not been established."[32] Table 5 contains a summary of the ILS visibility and DH requirements. However, it should be recognized that the widespread use of DGPS for precision approaches would probably lead to the definition of new landing minima criteria, such as the tunnel concept discussed below.

The microwave landing system (MLS), presently proposed in the International Civic Aviation Organization (ICAO) as the replacement system for ILS, is specified to have essentially the same sensor accuracies as CAT III ILS. Because DGPS also is now considered a serious replacement for ILS, the ILS sensor accuracy requirements are used here, where applicable, for specifying estimates

Table 5 Traditional categories of precision approach

Category	Visibility or runway visual range	Decision height
CAT I[a]	0.5 mile visibility or 2,400–1,800 ft	200 ft
CAT II[a]	1,200 ft	100 ft
CAT IIIa[b]	> 700 ft	DH < 100 ft
CAT IIIb[b]	150 < RVR < 700 ft	DH < 50 ft
CAT IIIc[b]	RVR < 150 ft	0

[a] Ref. 1.
[b] Ref. 32.

of DGPS accuracy requirements for the various categories of approach. The technique used is that of Ref. 13 where the requirements are derived from ICAO Annex 10.[8] In this technique, the sensor error components are categorized as is done in MLS.

Both ILS and MLS are systems that primarily measure the *angle* to the approaching aircraft; therefore, the errors translate into varying aircraft displacements depending upon the range to the aircraft from the ground antennas. On the other hand, GPS errors have no significant spatial variation through the approach path. This fundamental difference in the two types of systems makes it complex to translate the accuracy requirements established for ILS and MLS into similar ones that are applicable for DGPS.

In Annex 10, the MLS sensor error is broken down into three components.

1) Path following error (PFE) is "that portion of the guidance signal error which could cause aircraft displacement from the desired course line or desired glide path."

2) Path following noise (PFN) is "that portion of the guidance signal error which could cause aircraft displacement from the mean course line or mean glide path."

3) Control motion noise (CMN) is "that portion of the guidance signal error which causes control surface, wheel, and column motion and could affect aircraft attitude angle during coupled flight, but does not cause aircraft displacement from the desired course and/or glide path."

In terms of DGPS errors, PFE is interpreted as any errors that remain constant during an approach plus any slow varying errors (e.g., waypoint coordinate error + multipath at DGPS reference station). Path following noise, a component of PFE, is interpreted as an error that varies slowly during the approach (e.g., multipath at DGPS reference station), and CMN is interpreted as an error that varies fairly rapidly so such that the aircraft body may respond to it by attitude changes only (e.g., multipath at the aircraft that is reflected from the terrain). As derived from Annex 10 error tolerances, Table 6 contains estimates of the above three portions of errors as a function of approach category. In developing Table 6, PFE is derived from the RSS of ILS alignment and beam bend error tolerances; PFN is ILS beam bend error tolerance; and CMN is based on MLS (because

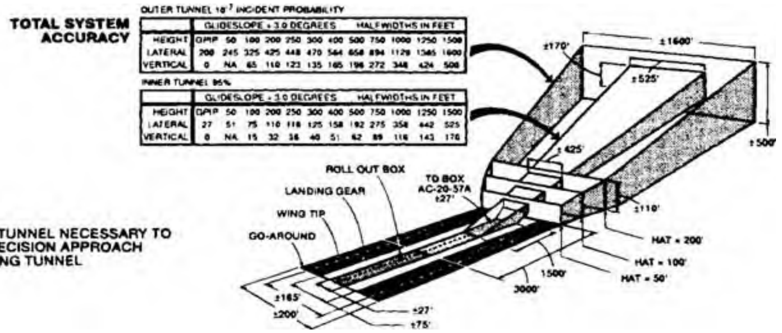
Table 6 Estimated precision approach 95% sensor accuracy requirements for DGPS at the decision height (9000-ft runway)

Approach category	Error direction	Path following error, ILS, m	Path following noise, ILS, m	Control motion noise, MLS, m
CAT I,				
DH = 200 ft	Lateral	16.0	13.3	4.3
	Vertical	4.1	2.9	1.2
CAT II,				
DH = 100 ft	Lateral	5.0	3.8	3.6
	Vertical	1.8	1.0	0.6
CAT III,				
DH = 50 ft	Lateral	4.0	3.5	3.2
	Vertical	0.6	0.5	0.3

there are no CMN requirements for ILS). Also, the Annex 10 alignment error is considered a 3σ error because it is an "adjust and maintain" tolerance.

In the ICAO Annex 10 method of assessing PFE and CMN, measured sensor errors are passed through filters. One filter is a second-order filter whose output represents PFE, and the other is a high-pass filter whose output represents CMN. The filter outputs are compared to the sensor accuracy requirements. The procedure and the filter parameters can be found in Ref. 8, and the results of their application to DGPS guidance errors measured during flight test can be found in Ref. 33.

The FAA's SOIT has developed a tunnel-in-space (tunnel) concept that focuses on total system error rather than sensor error when specifying RNP.[34] The tunnel emphasizes the continuous containment of total system error from the final approach fix through rollout. The lateral and vertical dimensions of the tunnel decrease with decreasing distance from the runway threshold. Unlike the conventional approach of emphasizing sensor errors, it gives credit to the users who attain small FTE and allows a tradeoff between sensor accuracy and FTE. The tunnel concept is defined by an inner and outer tunnel, as shown in Fig. 6.[34] The inner tunnel is a surface within which the guidance reference point on the aircraft must remain with a probability of 95%. The outer tunnel is an outer containment surface where any part of the aircraft must not penetrate. The maximum penetration probability is 1 × 10^{-7} for vertical or lateral penetration per approach because of navigation errors. The basic total system accuracy performance is based on the inner surface because it is directly measurable through flight testing. The outer tunnel performance is mainly evaluated by analysis and simulation. The definition of the tunnel surfaces are continuous; therefore, in principle, it could allow users to attain a range of landing minimums, depending upon equipage for RNP and the local obstacles about the approach path (e.g., towers and high terrain). Table 7 contains tunnel total system requirements at various DH points along the final approach path.[34] In contrast to the sensor errors shown in Table 6, there is no specification for 50-ft DH because, at this point, it is assumed that for a CAT III approach, vertical guidance is mainly provided by a radar altimeter.



TOTAL SYSTEM ACCURACY

OUTER TUNNEL 10^{-7} INCIDENT PROBABILITY

	GLIDESLOPE = 3.0 DEGREES						HALFWIDTHS IN FEET					
HEIGHT	GPIP	50	100	200	250	300	400	500	750	1000	1250	1500
LATERAL	200	245	325	425	448	470	564	658	894	1129	1365	1600
VERTICAL	0	NA	65	110	123	135	165	196	272	348	424	500

INNER TUNNEL 95%

	GLIDESLOPE = 3.0 DEGREES						HALFWIDTHS IN FEET					
HEIGHT	GPIP	50	100	200	250	300	400	500	750	1000	1250	1500
LATERAL	27	51	75	110	118	125	158	192	275	358	442	525
VERTICAL	0	NA	15	32	36	40	51	62	89	116	143	170

■ ONLY ONE TUNNEL NECESSARY TO DEFINE PRECISION APPROACH AND LANDING TUNNEL

Fig. 6 Definition of inner and outer surfaces of the RNP tunnel.[34]

Table 7 Precision approach total system accuracy requirements based on the tunnel concept[34]

Tunnel surface	Decision height, ft	Vertical 1/2 width, ft	Lateral 1/2 width, ft
Inner, 95%	200	32	110
	100	15	75
	50	NA	51
	Runway	NA	27
Outer, 10^{-7}	200	110	425
	100	65	325
	50	NA	245
	Runway	NA	200

Note: aircraft dimensions must be subtracted from outer tunnel limits.

2. *Integrity*

Two basic alternatives are discussed for ensuring signal integrity for precision approach. One approach is similar to the monitoring of the ILS or MLS. In ILS and MLS, the transmitted signal alignment error is monitored. If the monitor alarm limits are exceeded for some specified period of time the system is shut down or there is a transfer to a standby transmitter. In a similar manner, the DGPS data transmitted to the aircraft can be independently monitored on the ground to ensure that the transmitted differential data have integrity. As noted in the discussion on nonprecision approach, monitored differential data ensures the truncation of large correction errors, and provides, in a sense, monitored navigation signals. Based on an ICAO Annex 10 standard[8] for ILS signal monitoring (shift of mean course line or glide path angle), Table 8 contains estimated precision approach monitor protection levels for each category of approach. These are only estimated protection levels for errors in differential corrections because they are based on ILS. However, they could be conservative levels given that GPS has less unobservable errors than ILS with respect to errors that cannot be monitored on the ground, such as multipath at the aircraft.

It should be noted that in comparing the monitor limit values in Table 8 to the PFE accuracy requirements in Table 6, in some cases, the monitor limits are smaller than the PFE sensor accuracy (CAT I) or comparable to them. This is explained by recalling that the PFE is composed of both alignment and PFN, and PFN is not monitored because it is a local error at the aircraft position. For a ground-based landing system, the aircraft antenna is pointed toward the ground and moving toward the source of radiation in such a way that any multipath error would be of low frequency (course bend). This results in an aircraft displacement. With DGPS, the most likely source of low-frequency multipath would be at the ground reference station antenna. Therefore, this antenna must be sited very carefully.

Another approach is where integrity monitoring is performed onboard the aircraft. In this method, the current flight technical error (FTE) is subtracted from the outer tunnel surface (Table 7). The result is compared to the current estimate of sensor error. The sensor error estimate includes estimated errors in

Chapter 13

GPS Applications in General Aviation

Ralph Eschenbach*
Trimble Navigation, Sunnyvale, California 94088

THE Global Positioning System (GPS) will have a profound impact on general aviation. Many observers believe that by the end of the decade, GPS will play an important role in the three principal aspects of flying required to allow a flight between two airports without outside intervention or assistance: navigation, collision avoidance, and landing. We look at these three in more detail; however, first we look at the market served and current solutions to these problems.

I. Market Demographics

The term *general aviation* (GA) usually is applied to all noncommercial aircraft applications. This includes all private, corporate, and business aircraft, but excludes commercial airline aircraft.

A. Airplanes

The United States dominates the general aviation airplane market. More than 75% of the GA fleet is located in the United States. As shown in Table 1, there are about 265,000 registered aircraft in the United States, of which 212,000 are active.

About 60% of these aircraft are primarily for personal use, and 40% are for business use. In addition to these GA aircraft, there are about 6000 air carrier aircraft. In all, GA accounts for about 67% of all hours flown, 25% of all passengers flown, and 45% of all miles flown. General aviation does this while burning only about 7% of the fuel consumed.[1] It plays a big role in transportation and interstate commerce, and GPS, by enhancing safety and reducing costs, will make this role even more important.

B. Pilots

In 1990, there were about 702,000 active pilots. They are broken into the categories shown in Table 2.[2]

*Vice President, Navigation.

Table 1 Number of active general aviation aircraft in 1990 by type and primary use

Aircraft type	Active GA aircraft	Corporate	Business	Personal	Instructional	Aerial Application	Aerial observation	Other work	Commuter air carrier	Air taxi	Other	Inactive
Total all aircraft	212,229	10,906	35,496	120,636	19,889	6687	5302	1525	1242	6188	4,358	54,115
Piston, total	187,773	3933	33,863	113,429	18,603	5402	4,011	1041	643	3853	2,995	46,285
One-engine	165,073	1412	25,615	106,868	17,686	5152	3779	951	303	928	2,380	42,311
Two-engine	22,606	2521	8248	6559	915	234	228	90	284	2925	603	3,886
Other piston	94	0	0	3	3	16	4	0	56	0	12	88
Turboprop, total	5652	2861	847	262	38	220	23	16	466	640	280	759
Two-engine	5257	2856	834	224	38	44	22	13	439	547	240	655
Other turboprop	395	5	12	38	0	176	1	3	28	93	39	104
Turbojet, total	4374	3204	340	115	4	0	17	0	0	374	321	517
Two-engine	3950	2938	329	113	1	0	17	0	0	343	209	355
Other turbojet	425	266	11	3	3	0	0	0	0	31	112	161
Rotocraft, total	7397	863	393	1369	877	1,065	995	224	126	1,132	355	3025
Piston	3459	45	133	1174	798	723	412	65	2	0	108	2343
Turbine	3938	818	260	195	79	342	583	159	124	1,132	247	682
Other, total	7032	45	55	5459	367	0	256	245	7	190	408	3530

Source: FAA.
Note: Row and column summation may differ from printed totals because of estimation procedures, or because some active aircraft did not report use.

Table 2 Estimated active pilots

Students	128,663
Private	299,111
Commercial	149,666
Airline transport	107,732
Miscellaneous	17,487
Total	702,659
Flight instructor	63,775

Despite the general decline in the GA market, which has seen new aircraft sales drop from about 17,000 in 1978 to about 1200 in 1990,[3] the number of pilots and the number of hours flown has remained constant. During this time, the general proficiency of pilots has improved dramatically. Instrument ratings (which allow a pilot to fly in clouds) have increased as a percentage of active pilots from 38% in 1976 to 52% in 1990.[2] Here again the capabilities of GPS will greatly accelerate this trend.

During the last 20 years, GA safety has improved also. The accident rate has declined by 60%, and the fatal accident rate has been more than halved. This reflects both pilot proficiency and aircraft instrumentation. The GPS will greatly enhance the pilot's situation awareness, and this will improve the safety record even more.

C. Airports

The area of airports and their usage is where the GPS will have its most dramatic impact. Currently, about 17,500 airports are in use in the United States, and about 5200 of them are in public use.[2] Of these, only about 1100 (or 6%) have Instrument Landing Systems (ILS).[2] GPS will allow an all-weather approach to be made at virtually all of the airports. Thus, GPS may expand the airport landing capacity by an order of magnitude.

II. Existing Navigation and Landing Aids (Non-GPS)

People have been flying airplanes for many years before GPS arrived. How did they navigate? The history of navigation is a long one, and here we look only at the era of flight and, in particular, at radio navigation systems.

A. Nondirectional Beacons (NDB)

The first radio navigation systems were nondirectional beacons. For the most part, they are in the 200–400 kHz band, though AM broadcast stations from 550 to 1600 kHz can also be used as NDB. By using a pair of directional antennas, a receiver can be designed to display the relative bearing between the aircraft heading and the station. Such a receiver is called an automatic direction finder (ADF). These are still found in most cockpits today, but they are seldom used by the modern navigator. Automatic direction finders frequently are used today in less developed parts of the world.

By keeping the relative bearing zero degrees, a pilot can use the ADF to fly to the station. Note that in the presence of winds, this does not result in a direct path to the station, but it will get the pilot there nonetheless.

The major problem with this system is that the pilot cannot determine position unless the pilot crosses directly over the station. With the use of a compass, however, the pilot can determine the radial from an NDB, as shown in Fig. 1.

Because an NDB does not take winds into account, it is very difficult to use an NDB to fly a given radial to or from a station. This problem is solved by the advent of the vhf omnidirectional radio (VOR) system.

B. Very High Frequency Omnidirectional Radio

The VOR is the backbone of the current air traffic control (ATC) system. At present, there are some 1000 VORs in service. The VOR system allows a pilot to determine the radial from a station *independent* of the aircraft heading. Each VOR station transmits in the band from 108 to 118 MHz. A 9.96-kHz tone is FM modulated ± 480 Hz by a CW signal. The antenna is rotated to create an amplitude-modulated signal in the receiver as the antenna sweeps by. The FM is synchronized to the rotation so that there is zero phase difference between the FM and AM, when the antenna points due north. The phase angle between the AM and FM signals is the radial from the station.

Although this system does not yet determine aircraft position directly, two VORs can be used to get two radials, and the aircraft position is estimated to be at the intersection of these two lines. An alternate rule of thumb used by many pilots to determine the range from a single VOR is to fly perpendicular to the VOR and measure the time in minutes it takes to change the radial by one deg. The distance from the VOR in nautical miles (n.mi.) is equal to the pilot's speed in knots (kts) times the time in minutes to change 1 deg. Thus, if it takes 30 s to move 1 deg at 140 kts, then the pilot is 70 n.mi. from the station. A more

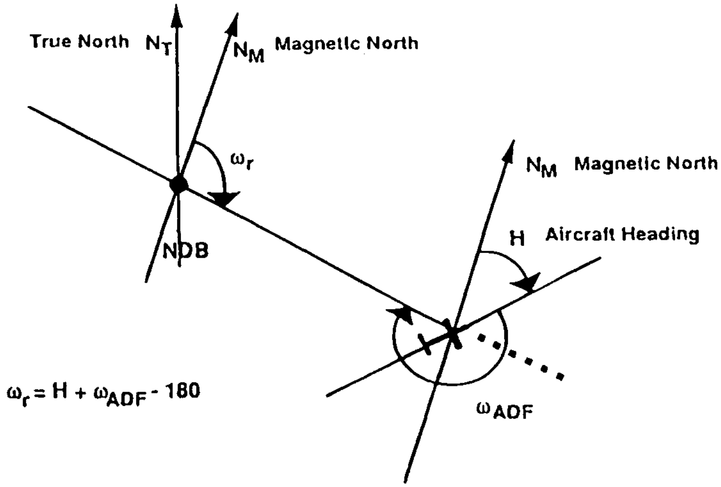


Fig. 1 Use of a nondirectional beacon to determine a radial.

accurate and quicker method to determine range from a station came about with the development of distance-measuring equipment (DME).

C. Distance-Measuring Equipment

Distance-measuring equipment is based on measuring the propagation time of a signal from the airplane to the DME station, and back. DME is collocated with VORs, thus enabling the pilot to get both range and radial from the same point to quickly determine position. DME operates in the band from 960–1215 MHz. DME is linked to VORs so that the VOR frequency is used to identify the DME for that station.

An interesting problem with DME is that of determining which reply from a given station is the reply triggered by the transmission from the pilot's aircraft. Remember that all airplanes using the same DME are transmitting on the same frequency. Each radio transmits a sequence of pulses at a random interval from 5/s to 150/s. The receiver then gates a sliding window after each pulse. When replies are received *repeatedly* in the same window, those replies must be in response to that radio's interrogations. Thus, the time shift of the window is a measure of the distance to the station.[4]

The VOR/DME system is the basis for most overland aircraft navigation used in the world today. It is also used as the core of nonprecision approaches used throughout the world. One problem with VOR/DME is that it is a "line-of-sight" system; thus, the maximum useful range is about 70–100 miles. If a mountain range comes between the aircraft and the station, the system will not work, and the range could decrease substantially. The long-range radio navigation (LORAN) system attempts to solve that problem.

D. Long-Range Radio Navigation

The LORAN system was originally developed as a marine system, and it is still used primarily on the coastlines of the Northern Hemisphere. A LORAN chain is made up of three to six transmitters linked together as a master and multiple secondaries. They all transmit a series of pulses with a common repetition rate called a group repetition interval (GRI). By measuring the time difference (TD) between the arrival of the master pulses, and a given secondary's pulses, a line of position (LOP) can be formed, which results in a hyperbola with the two transmitters as foci. LOPs from two master–secondary pairs allows a point position to be determined, as shown in Fig. 2. Ambiguous solutions can arise, as seen in Fig. 2. When this occurs, an additional LOP can be used to resolve the ambiguity. If an additional LOP is not available, some other information must be used for resolution. In the worst case, the equipment notifies the user of an ambiguous solution, and the user is asked to determine the correct solution.

There are currently 17 different chains (called GRIs) located around the world transmitting on 100 kHz. The GRI designates the group repetition interval of the pulses for that particular chain. The low frequency gives the system a range of about 1000 n.mi. For many years, only U.S. coastlines were covered, but recent expansion of the system closed the "midcontinent gap," and we now get good coverage in the contiguous states. Depending upon the geometry of the transmitters, LORAN has an absolute accuracy of about 0.25 n.mi. and a repeatable

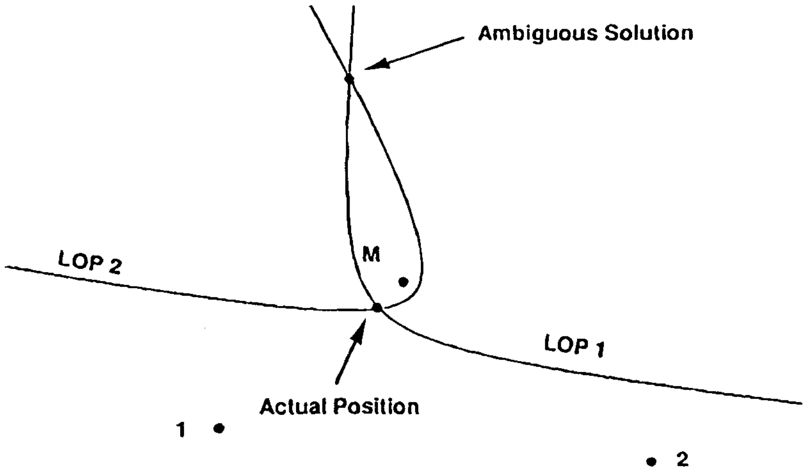


Fig. 2 LORAN is a hyperbolic system.

accuracy between 18 and 90 m.[5] Currently there are about 600,000 users worldwide.

One difficulty of the LORAN system is that it is susceptible to low-frequency noise. There are many man-made sources, including power lines, high-power Navy communication transmitters, street cars, and more, but the most detrimental to the GA usage is that caused by electrical storms. In extreme cases, this can cause a complete inability of the receiver to track the signal. Another limitation of LORAN is geographic. There is no LORAN coverage more that 1000 n.mi. from any coast over water, and there is no coverage in the southern hemisphere at all. The coverage problem is solved by Omega.

E. Omega

Omega is also a hyperbolic system, but provides worldwide coverage by transmitting at an even lower frequency than LORAN (on four frequencies from 10.2 to 13.4 kHz). It is currently the only radionavigation system certified for extended over water flights. Eight stations give worldwide coverage. Accuracy of the Omega system is usually assumed to be about 4 n.mi. It is currently estimated that there are about 27,000 users.[6]

F. Approaches

Approaches are categorized as either precision or nonprecision. A nonprecision approach is one that does not provide glide path guidance (see FAA Document 7110.65G). Precision approaches are further categorized as Category I, II, or III. Table 3 shows the accuracies required for each.

Whereas VORs are usually named with a three letter identifier (e.g., SJC, SFO), approaches are usually named for the type of navigational aid used.

Table 3 Accuracy requirements for different types of approaches

Precision approach type	Accuracy, m
Category I	
Horizontal	16.5
Vertical	3.4
Category II	
Horizontal	6.5
Vertical	1.6
Categoty III	
Horizontal	4.1
Vertical	0.5

Note: All values are 95% limits.

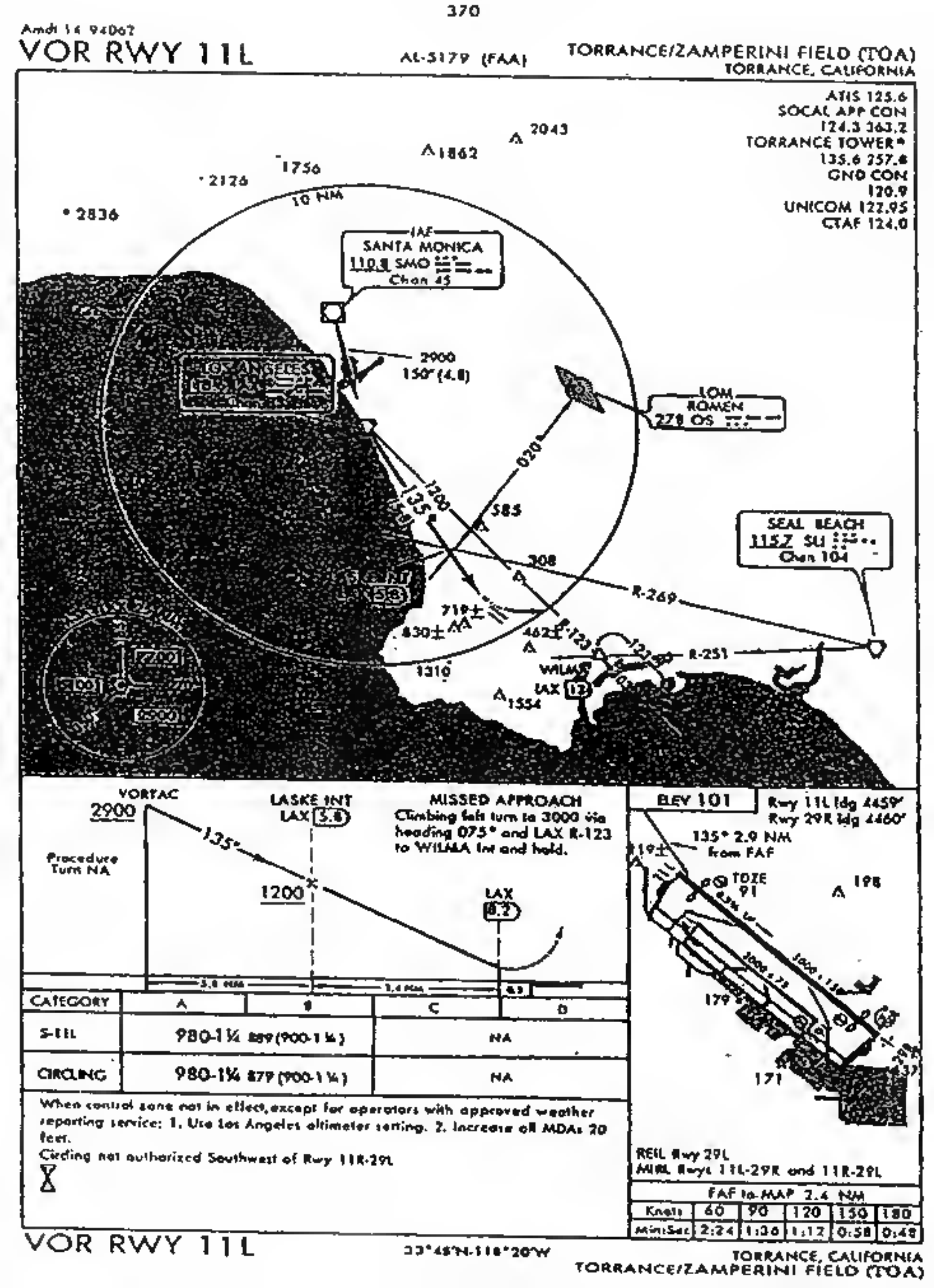
1. *Nonprecision*

There are four different nonprecision approaches—NDB, VOR, VOR/DME, and localizer back course. In all cases, the equipment required for the approach is included in the name. Thus, to do a VOR/DME approach, both a VOR and a DME must be in the airplane and in working condition. In a nonprecision approach, there is no vertical guidance from the navigation aid. Instead, the pilot uses the altimeter and descends in steps to specific minimum altitudes. In the example shown in Fig. 3, the pilot would maintain 2900 ft until crossing the Los Angeles International Airport (LAX) VOR, then he or she must descend to 1200 ft and remain at that altitude until LASKE intersection. After LASKE intersection, the pilot must descend to 980 ft until reaching the missed approach point (MAP).

Notice that one VOR (LAX) is used for guidance, and a second, SLI, is used as a cross-radial for position determination. In this example, it is used for both the final approach fix (FAF), and missed approach hold point (MAHP). Also, in a nonprecision approach, the MAP is frequently determined by time and velocity. This is seen in the table at the lower right of the approach plate. In a nonprecision approach, after the FAF (here LASKE INT), the pilot descends to the minimum descent altitude (MDA) (here 980 ft MSL). This altitude is held until the time has elapsed from the FAF to the MAP (here 1 min 12 s for a speed of 120 kts). If the field is not in sight at that time, a missed approach is executed. The sole altitude reference for this approach is a baro-altimeter. The GPS altitude cannot be used for the approach.

2. *Precision Approaches*

Precision approaches, and ILS in particular, are the standard civil landing system used in the United States and abroad. Whereas nonprecision approaches use the altimeter for vertical guidance, precision approaches use a radio signal, and the altimeter is used only to determine the decision height (DH). This is the altitude at which a decision must be made whether to execute a missed approach or not. A typical ILS approach is shown in Fig. 4.



CATEGORY	A	B	C	D
S-11L	980-1¼ 889 (900-1¼)		NA	
CIRCLING	980-1¼ 879 (900-1¼)		NA	

When control zone not in effect, except for operators with approved weather reporting service: 1. Use Los Angeles altimeter setting. 2. Increase all MDAs 20 feet.

Circling not authorized Southwest of Rwy 11R-29L

FAF to MAP 2.4 NM					
Knots	60	90	120	150	180
Min:Sec	2:24	1:36	1:12	0:58	0:48

Fig. 3 A VHF omnidirectional radio system approach in Torrance, California.

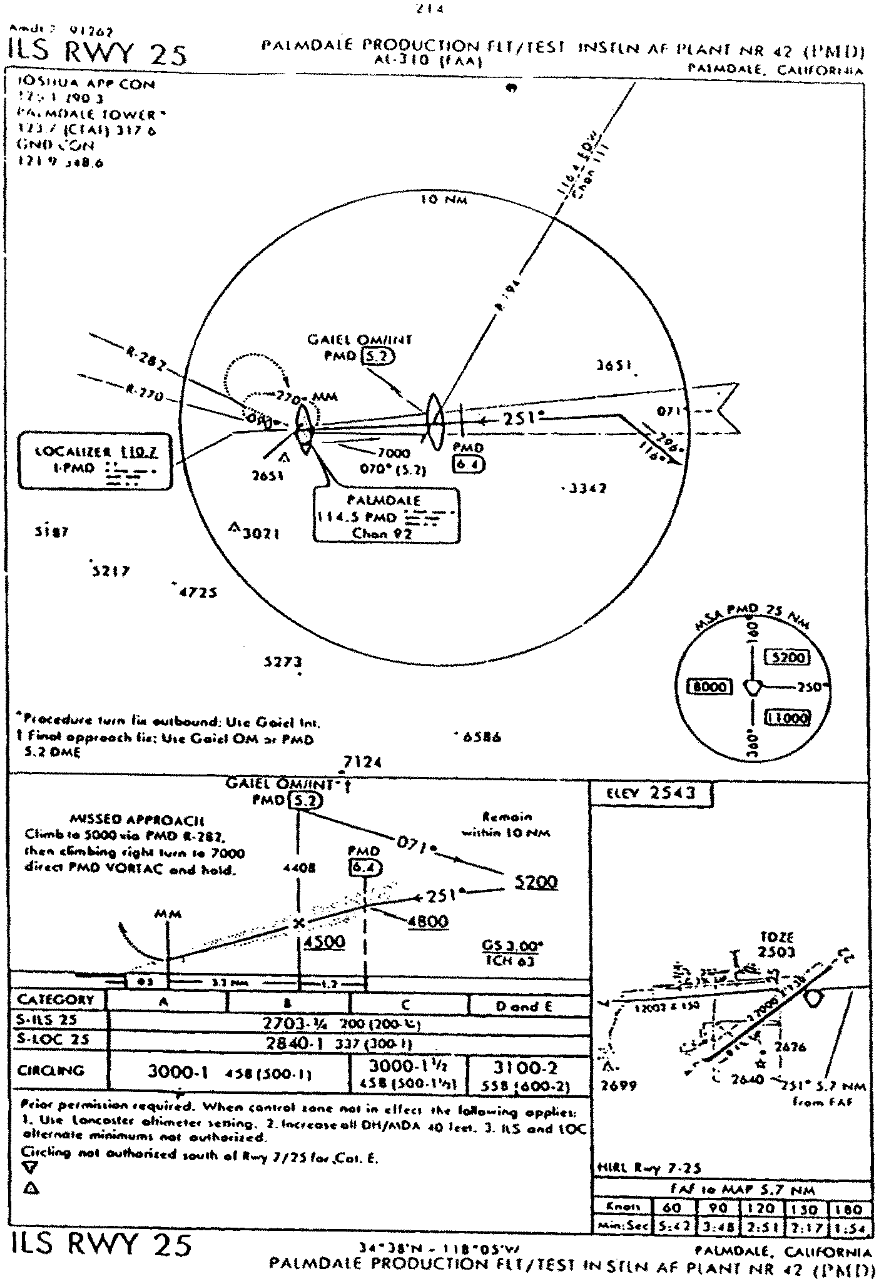

214
Amdt 2 91262
ILS RWY 25
PALMDALE PRODUCTION FLT/TEST INSTLN AF PLANT NR 42 (PMD)
AL-310 (FAA)
PALMDALE, CALIFORNIA
JOSHUA APP CON
124.1 290.3
PALMDALE TOWER*
123.7 (CTAF) 317.6
GND CON
121.9 348.6
116.4 EDW
Chan 111
10 NM
R-194
GAIEL OM/INT
PMD 5.2
R-282
R-270
270° MM
071°
251°
3651
LOCALIZER 110.7
I-PMD
2651
7000
070° (5.2)
PMD
6.4
296°
116°
3342
PALMDALE
114.5 PMD
Chan 92
5187
3021
5217
4725
5273
MSA PMD 25 NM
160°
5200
8000
250°
11000
340°
*Procedure turn fix outbound: Use Gaiel Int.
† Final approach fix: Use Gaiel OM or PMD 5.2 DME
6586
7124
GAIEL OM/INT* †
PMD 5.2
MISSED APPROACH
Climb to 5000 via PMD R-282, then climbing right turn to 7000 direct PMD VORTAC and hold.
Remain within 10 NM
4408
PMD
6.4
071°
5200
251°
4800
MM
4500
GS 3.00°
TCH 63
0.3
3.3 NM
1.2
CATEGORY | A | B | C | D and E
S-ILS 25 | 2703-¾ 200 (200-¾)
S-LOC 25 | 2840-1 337 (300-1)
CIRCLING | 3000-1 458 (500-1) | 3000-1½ 458 (500-1½) | 3100-2 558 (600-2)
Prior permission required. When control zone not in effect the following applies: 1. Use Lancaster altimeter setting. 2. Increase all DH/MDA 40 feet. 3. ILS and LOC alternate minimums not authorized.
Circling not authorized south of Rwy 7/25 for Cat. E.
ELEV 2543
TDZE 2503
12002 X 150
2676
2699
2640
251° 5.7 NM from FAF
HIRL Rwy 7-25
FAF to MAP 5.7 NM
Knots | 60 | 90 | 120 | 150 | 180
Min:Sec | 5:42 | 3:48 | 2:51 | 2:17 | 1:54
ILS RWY 25
34°38'N - 118°05'W
PALMDALE, CALIFORNIA
PALMDALE PRODUCTION FLT/TEST INSTLN AF PLANT NR 42 (PMD)

Fig. 4 Typical instrument landing system approach.

There are several things to note about this type of approach. From about 6 miles out, the vertical guidance is provided by the ILS. When the pilot reaches 2703 ft MSL (200 ft above the ground), a decision must be made. If the runway is not clearly visible, a missed approach must be executed.

The accuracy of ILS is sufficient for Category I, II, and III approaches, however it has limitations including siting, cost, frequency allocation, and performance. These are some of the reasons that there are only 1100 ILS approaches in over 5000 public airports. The siting problem occurs because the ILS requires a long, straight approach that is clear of obstructions. The microwave landing system (MLS) attempts to overcome some of these problems.

3. Microwave Landing System

The MLS is being developed by DOT, DOD, and NASA to replace the ILS. Because this system allows curved and steep approaches, it will allow closer spacing of aircraft on the same approach, as well as allowing approaches in more difficult terrain. The MLS operates by transmitting multiple signals to determine azimuth, elevation angle, and range to the end of the runway. The angles are determined by using scanning beams operating in the 5.25 GHz band. Range measurements are made in the DME band. The FAA has only recently begun to phase in the MLS. The transition will be slow. Some segments of the aeronautics industry doubt that MLS will ever be fully implemented because of the potential of satellite systems like GPS and GLONASS. The Air Transport Association has stated that its members believe that satellite systems will be able to demonstrate Category I approach capability by 1994, and Category III by 1997[5] (see also Chapter 12, this volume).

III. Requirements for GPS in General Aviation

To understand GPS in aviation, we must understand the certification environment. Specifically, we must understand Technical Standards Orders (TSO), Supplemental Type Certifications (STCs), and Form 337s. A TSO is a document put out by the FAA that outlines the specifications a piece of equipment must meet in order to comply with the FAA requirements. These include, but are not limited to, environmental, performance, user interface, and system interface requirements. An STC is a document that describes a typical installation in a particular airplane type. An avionics installer would use this document to comply with the installation requirements. If a TSO and an STC are not available for a given product, an installer may use a Form 337, but this is very difficult, because it requires local FAA approval.

The first receivers specifically designed for the general aviation market became available in 1990. These units were very similar to the LORANs available at that time in that they contained complete databases, and performed such typical area navigation functions as great circle range and bearing, ground speed, ETA, ETE, and much more. With GPS, however, the pilot got improved accuracy, quicker response to dynamics, much more accurate velocity measurements, and no geographic gaps in the coverage. These first units were installed with FAA Form 337 approvals. The STCs were first received about a year later. This allowed more general installations, and allowed the units to be connected to other aircraft systems. The first TSOs were accomplished with the help of other approved systems in a multisensor application. GPS/LORAN and GPS/Omega TSOs were received in the spring of 1992. On December 10, 1992, a GPS TSO was issued under the number TSO-C129. The first equipment was approved under this TSO in August, 1993, in Category AII. Category AII applies to a stand-alone unit operating in the Terminal and Enroute phases of flight. The next approvals for GPS are for the "approach overlay" program. This is described as TSO-C129 Category AI and allows GPS approaches to be flown "over" existing nonprecision approaches. Although GPS is more accurate than the other systems, the minimum descent altitude will not be improved at this time.

A. Dynamics

In GA, the dynamics are relative benign. Typical accelerations are in the range of 1–3 *g* with acrobatic applications up to 4–6 *g*. This presents little challenge for most GPS receivers. Roll rates are typically around 10 deg/s, with maximum rates of 60 deg/s. In nonacrobatic applications, these rates do not continue past 60–70 deg of bank. A standard rate turn is 3 deg/s. Once again, this is not difficult for a normal GPS receiver.

B. Functionality

The primary function of a GPS receiver in a cockpit is to enhance the pilot's position awareness and to provide information for navigational guidance. For a pilot, position awareness is in terms of position relative to a known ground navigation aid (such as VORs, NDBs, or airports) A pilot will not say, "I'm at latitude 37 23.6 N and longitude 122 2.3W," but will instead say "I'm on the 268 degree radial, 4.6 miles from San Jose VOR."

In terms of track guidance, the pilot is used to using a course deviation indicator (CDI). This is an analog instrument that shows the angular error between the desired track and the current position as seen from the destination. In most avionics GPS receivers, the CDI is used to display the cross track error (XTE). The usual scale is about 1 nm per dot with a five-dot range right and left.

C. Accuracy

Avionics accuracy requirements are different for each of the different phases of flight. These are usually separated into *en route, terminal, approach,* and *landing.* Table 3 shows the accuracy requirements for landing, and Table 4a shows those for the other phases of flight.

Table 4a Navigation accuracy requirements

Phase of flight	Accuracy
En route	2 n.mi.
Terminal	1 n.mi.
Approach	0.3 n.mi.

Table 4b GPS accuracy levels

Standard positioning service (SPS)	100 m
SPS without selective availability (SA)	25 m
Differential GPS	2–5 m
Kinematic carrier tracking	10 cm

The GPS has four different levels of accuracy. These are shown in Figure 4b. If we compare these levels with the requirements of the different phases of flight, we see that GPS standard positioning service (SPS) accuracies are sufficient for all phases of flight except landing, and for landing, we require differential GPS (DGPS). An interesting consequence of this is that selective availability (SA) is of no consequence for normal avionics applications. For en route and terminal applications, the expected level of SA has an insignificant effect, and for landings, differential GPS must be used in any case.

D. Availability, Reliability, and Integrity

For landing applications in particular, three other critical items must be addressed: availability, reliability, and integrity.

1. GPS Availability

Will the system be made available to the worldwide flight user community? The United States has stated that GPS will be made available. After KAL flight 007 was shot down over Russia, Larry Speakes, then deputy press secretary to President Reagan, stated that "the President has determined that the United States is prepared to make available to civilian aircraft the facilities of its Global Positioning System. . . ."[7] In addition, the United States has stated to the ICAO that the system will be available to users without charge for a minimum of 10 years.

Availability is not only a political question, but also a technical one. How many outages and of what duration will there be? The outages could be either planned (maintenance or improvements) or unplanned (equipment failures or lack of satellite visibility). For landing systems, for instance, a total system availability of 95% is expected. This must include the ground system, the user equipment, the datalink, and the satellite system.

2. GPS Reliability

How reliable will the system be? Most simulations show that GPS by itself would not meet the reliability requirements. Under specific scenarios of satellite failures, there would not be enough satellites in view to *guarantee* a reliable solution. Most solutions to this usually involve additional satellites. This includes GLONASS (the Russian equivalent of GPS), INMARSAT satellites, and pseudosatellites (pseudolites). Pseudolites are ground transmitters that transmit on the GPS frequency, and the user can both range and get differential corrections from this link. Most recently, the FAA is proposing to use synchronous satellites that have ranging capability and, thus, increase both reliability and integrity (see Chapter 3, this volume).

3. *GPS Integrity*

How can we be certain that if an answer is displayed it is correct? The requirement here is both the ability to detect errors and the ability to report them to the pilot in a timely manner. If there were always five or more satellites available, and good geometry, this could be solved by receiver autonomous integrity monitoring (RAIM). In general, this is the ability of the receiver to determine when the solution can be trusted for the intended application. For the landing phase of flight, GPS alone cannot deliver the necessary integrity. About an additional 12 satellites would be necessary to meet the integrity requirement.

Another proposed solution to this problem is the use of a GPS integrity channel (GIC) (see Chapter 4, this volume). In this solution, ground monitors are used in combination with a communication link to detect and communicate a problem with the system to the pilot within two seconds. The assumed communication channel would be a number (1–4) of geostationary satellites that could provide ranging information, as well.

In summary, solutions exist that will allow GPS to be used as a precision landing system to open up virtually all the world's airports to all weather landing capability.[8] It is not clear which will be the eventual system of choice, but at this time the systems appear as shown in Table 5.

Once a communication link is required for the DGPS landing systems, two other important applications become available—tracking, and clearance delivery. A cooperative aircraft can transmit its position so that a base station, be that a private base station or an ATC center, can monitor the progress of the aircraft in flight without radar or voice contact. In addition, clearances, weather, or other traffic information could be transmitted to the pilot for improved situation awareness.

IV. Pilot Interface

Because of other cockpit demands, it is mandatory that the receivers be designed to facilitate the flow of information from the pilot to the navigation system and return. Both input and output have evolved substantially over the years.

A. Input

In almost all of the earlier navigation systems, the only input required by the pilot was to tune the receiver to a particular frequency. With the advent of long-range systems, LORAN, Omega, and now GPS, it became necessary to input destinations, and routes. In some of the earlier systems, this was accomplished with complete alphanumeric keypads. This was especially true in high-end sys-

Table 5 Landing type systems

Landing type	System
Nonprecision	SPS
Category I	Differential GPS
Category II, III	Real-time kinematic with pseudolite differential GPS

tems where panel space was available. In GA, however, the most common method now in use is done with two concentric knobs. One controls the cursor position, and the other scrolls the alphabet and the numerals. This method compromises panel space, and flexibility. Software has eased this transition by only allowing the letters which are possible to show and by completing the spelling with only what is available. For example, when searching for Albuquerque, only the first four letters need to be entered, and when going from "Alba" to "Albu," there are only three steps: "e," "i," and "u." The others are not possible combinations.

B. Output

In older navigation systems, the only output was the CDI. This output has been retained in newer systems, but much additional information is presented in both alphanumeric and graphic form. Information such as Desired Course, Actual Track, Range to Destination, Speed over the Ground, Time Enroute, Time of Arrival, and Cross Track Error are all available on most GPS receivers today. In addition, databases in the receiver contain information about the Airports, VORs, NDBs, and Intersections. Airport information such as City Name, Airport Name, Runway lengths, Field elevation, and Fuel availability are all available at the touch of a button. Frequencies such as ATIS, Tower, Ground, CTAF, and Unicom are also readily available.

Graphic information is also becoming available in the cockpit. The most common of these is the moving map. In the most simple form, this is a plan view of the surface of the Earth with relevant aviation data presented on it and the user is at the center of the picture. Typically, VORs and airports are shown. In some moving maps, the airways are shown, as well. Here we can also have access to the airport information listed above. Because of the three-dimensional aspect of GPS, we will soon see "tunnel in the sky" presentations to guide the pilot. This will give the pilot much more ability to guide the plane than the current dual needle system.

V. GPS Hardware and Integration

A. Installation Considerations

1. Antenna Siting

The basic tenet of GPS antenna placement is that the antenna must have a clear view of the sky. Usually the only thing that could shadow the antenna is the tail, and thus, the antenna is usually placed over the cockpit well forward of the tail structure. Shadowing by the wings and fuselage will also occur during turns, but nothing can be done about this short of complex dual antenna structures. Again, the software can help by using sophisticated algorithms for reaquisition after loss of lock. One of the better algorithms is to use a Doppler predictor based on current position and velocity. This requires at least a three-satellite solution. If a new satellite can be reaquired before an old one is lost, this assumption can be maintained.

2. *VHF Communications (Comm) Interference*

Because of an unfortunate choice of frequencies, the 12th and 13th harmonics of the Aircraft VHF Communication Band (118–136 MHz) lie directly in the GPS band (1575.42 ± 1 MHz). The 12 communication channels between 121.125–121.250 and 131.225–131.350 (25 MHz spacing) produce harmonics directly in the GPS band. Because they are directly in the band, there is nothing that can be done in the GPS receiver to filter them out. The only option is to filter them out before the comm antenna, and hope that there is no leakage directly out of the comm transceiver. Because the comm radios were there first, the STC for TSO-C129 requires that a test be performed to see if the particular installation causes interference with the GPS, and if it does, a filter is recommended to filter the comm radio output. If this problem cannot be solved, an IFR installation is not allowed.

B. Number of Channels

The problem above (loss of signals during turns) is simplified and minimized by having more channels. Thus, the adage that "more is better." Because there are rarely more than nine satellites in view at once, more than nine channels are of little value. With a sequencing receiver, this problem is exacerbated. Thus, for good performance in turns, a minimum of four channels is necessary, while six or more improves performance.

Another place where more channels helps is in *time to first fix* (TTFF). Because it usually takes about 5 min from power turn on until navigation information is necessary, TTFF is not an important issue in aviation.

C. Cockpit Equipment

As of this writing (1994) there are three types of GPS equipment available to the GA pilot. These are handheld, panel mount, and Dzus mount. All three types have gained rapid market acceptance as the constellation has filled. The market segments are described below, but more specific information can be found in Ref. 9.

D. Hand-held

The hand-held receiver is the lowest priced way to get GPS capability. This is also the best solution for the nonowner pilot. There are two entries into this market—the Garmin 55 AVD and the Trimble Flightmate. Both of these units have avionics databases containing the location of most airports and VORs. They are both battery powered and run for about 4 h on two AA cells. Inside an aircraft, they both work better with a remote antenna. A remote antenna can be mounted easily on the windshield for better satellite visibility. They both have a street price of about $1100. The units differ in the user interface and the display. These units have an RS-232 output, which can be connected to computers for data recording and processing.

E. Panel Mounts

This market has several entries. The principal products are made by Garmin, Trimble, King, Arnav, Narco, and IIMorrow. All have databases with complete information about airports, VORs, NDBs, and intersections. In addition, typically 100 user waypoints can be added. Output can drive moving maps, CDIs, HSIs, and more. Some units interface to air data computers and can automatically compute winds aloft, density altitude, and true airspeed. Interfaces to fuel sensors allow computation and display fuel consumption rates, fuel on board, fuel consumed, and fuel remaining at destination. Warnings are provided when fuel reserves are too low.

Although many similarities exist, there are some substantial differences. The number of GPS channels varies from Garmin with 1, to Trimble with 6, to Arnav with options from 5 to 12. The low end of this market uses LCD with most of the products using LEDs. King has a CRT display that gives added flexibility to the user interface.

Several of the high-end units interface to PCs, which allows for the ability to upload and download waypoints, flight plans, search patterns, and more. Commercial flight-planning software can now download flight plans directly into the unit or into a datacard that can be carried to the plane and loaded into the receiver.

F. Dzus Mount

The Dzus mount market, as the panel mount market, describes the installation method. Dzus refers to a 5-1/4 in.-wide package that usually mounts in the console between the two pilots on larger aircraft. This market is made up of the high-end twins, jets, and helicopters.

The primary participants in this market are Global, Trimble and Universal Navigation. This market requires FAA certification, usually in the form of a TSO. At present, GPS can only be TSOed in combination with other such systems as LORAN or Omega. Trimble is the only manufacturer that supplies certified equipment combining both GPS/LORAN and GPS/Omega. Interfacing to other aircraft systems is mandatory in this market. Roll steering and ARINC 429 interfaces are common. The digital interface protocol, ARINC 429, is used to communicate between different pieces of avionics equipment. Roll steering is an analog output that commands a roll angle to an autopilot. Roll in degrees is computed using the following formula:

$$\mathrm{ROLL} = -(0.00281*\mathrm{XTE} + 0.002*\mathrm{GS}*\mathrm{TKE})$$

where ROLL is bank angle in deg; XTE is cross track error in ft; GS is ground speed in ft/s; and TKE is track angle error in deg.

VI. Differential GPS

As pointed out earlier, the only way approach accuracies can be achieved is through the use of differential GPS. The three components of a DGPS system are the Reference Station, the Communication Link, and the Airborne Receiver. For Special Category I approaches (SCAT I), these components are defined in RTCA Document DO-217.[10]

A. Operational Characteristics

To fly a SCAT I, an ATC clearance is required. Obviously, SCAT I compliant equipment must be in the aircraft and on the ground. A DGPS status indicator must indicate that the unit is receiving and using differential corrections. Upon activation of the approach, both vertical and horizontal guidance will be based upon angular deviation from the desired course. Once the approach is selected, detection of failures or losses of integrity will be annunciated by flags in the guidance indicators. Once a flag is present, the pilot must not continue the specified approach under instrument meteorological conditions (IMC) using DGPS. If a missed approach is initiated before crossing the threshold waypoint, guidance will be given to that waypoint, and then the pilot is expected to execute the published missed approach.

It should be noted that this initial use of DGPS for precision instrument approaches will be "supplemental" only. That is, when operating under instrument conditions, use of DGPS can only be done if other appropriate landing systems are available.[11]

B. Ground Stations

The ground station is made up of four components: a DGPS reference receiver, a data-processing function, a DGPS signal integrity monitoring function, and a data transmitter. The reference receiver must compute pseudorange corrections with an accuracy of better than 1.1 m rms. Over a 2.5-min approach, it must have a failure probability (attributable to hardware failure or integrity alarm) of less than 3.8×10^{-5}. The integrity of the system will be designed to meet the requirement that the probability that any part of an aircraft penetrates the outer tunnel without warning will be less than 1 in 10^7 approaches.[11]

The data-processing requirements will be such as to compute and format the differential correction messages in the proper manner. A user differential range error (UDRE) is also computed as a measure of the pseudorange corrections generated by the reference station. The confidence in this limit shall be at least 99.5%.

The integrity monitoring function must be completely independent from the rest of the ground equipment. It must monitor the integrity of all the data generated by the ground equipment before it gets transmitted over the data link. If the monitor itself fails, the data link must be shut down immediately. In all cases, a failure must be communicated to the pilot within 3 s.

There are currently two proposals for the data link. One uses the VHF aircraft navigation (VOR) band, and the other uses Mode S. Reference 11 contains complete implementation details of these two approaches. In either case, the frequencies will be selected automatically once an approach has been initiated by the pilot. The pilot will not be able to select the wrong frequency, as is now possible with an ILS approach.

C. Airborne Equipment Features

The primary functions of the airborne equipment are to receive the differential correction, receive the GPS signals, compute the corrected GPS position and navigation solution, and manage the navigation database.

To ensure that position accuracy requirements are satisfied, navigational information used for display must be updated at a 5-Hz rate or more. In addition, the latency must be 0.2 s or less. Accuracy and alarms will be determined by the tunnel concept. This involves both an inner and an outer tunnel. The total system accuracy must be within the outer tunnel always (less than 1×10^{-7} incident probability) and within the inner tunnel 95% of the time.[11] Integrity warnings must have a latency of no more than 3 s. This is so that the overall system can maintain a 6-s warning of any malfunction to the pilot because the ground segment can also have a 3-s latency.

The navigation database also requires some modifications from the usual avionics database. In addition to some new waypoints, the resolution must be increased to 0.0001 min for latitude and longitude and 0.1 ft in altitude. The equipment will at a minimum store the *glidepath intercept waypoint* (GPIWP), the *threshold crossing waypoint* (TCWP), and the *threshold crossing height* (TCH). The GPIWP and the TCWP are used to define the *final approach segment* (FAS), and the TCH is used to define the containment tunnels.

To get complete system integrity, pilot error must also be reduced as much as possible. Pilots will not be allowed to manipulate individual waypoints, as is now possible for en route navigation. When an approach is selected by the pilot, all the appropriate waypoints will be automatically concatenated into a "route" to be flown by the pilot. The waypoints will be sequenced automatically, and missed approach guidance will be given.

VII. Integrated Systems

Until *initial operational capability* (IOC) was declared in December 1993, the availability of GPS was at times so poor that GPS could not be relied upon for a navigation system. Primarily for this reason, GPS was integrated with other systems to provide the reliability and availability needed for commercial navigation. These systems achieved credibility through the Multisensor TSO C-115a. In general, this TSO requires that if the two independent sensors disagree, GPS must be ignored. Also, because of integrity, the two systems should be kept as independent as possible. This requirement eliminated the option of integrating pseudorange measurements from different systems into one navigation solution.

A. GPS LORAN

This combination provided TSOed capability in the domestic en route structure. This allowed GPS to be used by GA pilots about 2 years before it would otherwise have been available. The main drawback to this system was LORAN's susceptibility to electromagnetic interference. Just when navigation becomes very important, when the weather is poor, LORAN would become unavailable.

B. GPS/Omega

GPS was first combined with Omega in 1991. This allowed the use of GPS for worldwide en route navigation. By adding a GPS sensor to an Omega navigation system, a seamless path was provided to the pilot to get GPS accuracies without additional training.

Now that GPS has achieved operational status, these integrated combinations will be less important to the aviation community.

VIII. Future Implementations

We have only begun to tap the potential that GPS brings to aviation. If the last decade is any indication, the next decade will see spectacular results. Navigation, collision avoidance, and landing systems will, of course, be improved, but there will be completely new applications, as well. A very exciting area is in the use of GPS for an attitude and heading reference system (AHRS).[10]

A. Attitude and Heading Reference System

Because we can measure differential phase between two antennas from a single satellite to an accuracy of about 1 mm, by tracking four or more satellites simultaneously, we can determine the relative position of two antennas to a few millimeters. If two antennas are placed about 2 m apart, we can then determine the vector between them with an angular resolution of 1 mrad. By adding another antenna perpendicular to the first, we can then determine the complete three-dimensional attitude of a body. If we now mount these three antennas on the two wings and tail of an airplane, we can use the system as an AHRS. This will allow the complete determination of the attitude without a gyroscope. At present, the update rates are too slow to replace the gyro in auto pilot applications, but a gyro replacement will certainly come in this decade. For a more detailed treatment of this, see Chapter 19, this volume.

B. Approach Certification

By using the tracking capabilities of GPS, approach certification and landing pattern determination will be made substantially easier. By tracking an airplane during a normal visual flight rules (VFR) approach a safe landing pattern can be determined. At a later time, that exact path could be uploaded to the pilot for an instrument flight rules (IFR) approach.

By using a computer to add the flight technical error (the error induced by the man–machine combination) to the desired approach pattern, a box could be flown in VFR that would be the limits of expected errors from the desired path. Again, in VFR, we could quickly determine if the approach with the expected errors was a safe one.

We can now imagine that the equivalent of an automated terminal information system (ATIS) would upload the current approach in use so that the pilot would only need to acknowledge the approach in use. This would eliminate the cumbersome job of waypoint entry during the most stressful part of an IFR flight.

C. Collision Avoidance

Another area where GPS will be exploited is the area of collision avoidance. Passive Traffic Alert and Collision Avoidance Systems (Passive TCAS) rely heavily on knowledge of position. GPS could provide this information very easily.

An even better solution involves broadcasting an exact time-tagged position. This could be done with a Mode S transponder transmitter. A cooperating aircraft could then receive these transmissions, and compute the projected paths of both the "own" and "other" aircraft, and warn the pilot or recommend an evasive maneuver if a collision were imminent. This could reduce the need for en route traffic advisories.

We could envision a modular system, where the simplest block is a low-cost GPS combined with a uhf transmitter. This combination would transmit the exact position, velocity, time, and ID of the host aircraft about once a second. This would have to be low cost because it would have to be mandated and carried on all aircraft to be effective. These transmissions could be listened to by the ATC system, and aircraft could be presented on a screen in a manner similar to that currently provided by radars. By phasing out radars, huge maintenance costs could be eliminated. Two additional modules would enhance this system: a database and display for navigation, and a receiver and display for collision avoidance (see also Chapter 12, this volume).

The Navigation Module (NM) would use the GPS signals to drive a navigation management system (NMS) similar to that provided by the current GPS receivers. A database containing waypoints, frequencies, airport information (runways, elevations, location), airways, minimum en route altitudes (MEA), etc., would eliminate the need for charts and their cumbersome usage in the cockpit.

The collision avoidance module (CAM) would receive the transmissions from other aircraft in the area, and would compute and display the relative position and velocity and, thus, the collision threat possibility. In a similar manner, transmitter modules could be placed at or near tall obstacles. The CAM would then display their positions, as well as those of nearby aircraft.

As can be seen, GPS will play a very important role in the three principal aspects of flight: navigation, landing, and collision avoidance.

D. Autonomous Flight

With the systems described above (navigation, collision avoidance and landing) we can easily imagine what might be called autonomous flight. In much the same way a person gets in a car and safely drives to a destination without having to file a "drive plan" and being in constant communication with a "Land Traffic Control Center," we can now imagine a person being able to get in a plane and safely flying to a destination without having to file a "flight plan" and being in constant control of an "Air Traffic Control Center". The ATC role would become one more of coordination than of control and separation.

Because of the tracking capabilities of GPS, the ATC will know the position and velocity of all aircraft in flight. This will allow them to monitor congested routes and airports and recommend alternates when necessary.

IX. Summary

As has been shown, GPS has already made a profound impact on general aviation. GPS has brought precise navigation to much of the world where little or no capability existed before. Collision avoidance and landing systems will soon be using GPS. When all three are in common usage, the impact will be even greater allowing autonomous flight for many applications. Even so, we've only just begun.

References

[1] Anon., "AOPA 1992 Aviation Fact Card," AOPA, 421 Aviation Way, Frederick, MD, 21701, 1992.

[2] Anon., *General Aviation Statistical Databook,* General Aviation Manufacturers Association, 1400 K Street NW, Suite 801, Washington, DC, 1992.

[3] Anon., "1991 Avionics Retrofit Market Analysis," Aircraft Electronics Association.

[4] Kayton, M., and Fried, W. R., *Aviation Navigation System,* John Wiley, New York, 1969, pp. 181–192.

[5] Langley, R., "The Federal Radionavigation Plan," *GPS World,* March 1992, pp. 50–53.

[6] Anon., "1990 Federal Navigation Plan," Copies available from NTIS, 5285 Port Royal Road, Springfield, MA 22161 as Document DOT-VNTSC-RSPA-90-3/DOD-4650.4.

[7] Montgomery, H., "Uncommon Ground," *GPS World,* June 1992, pp. 16–19.

[8] Lechner, W., "The Potential of Global Satellite Systems for Precision Aircraft Navigation," *GPS World,* June 1992, pp. 40, 41.

[9] Connes, K., *The Loran, GPS, & NAV/COMM Guide,* Butterfield Press, 1992, pp. 98–99.

[10] Cohen, C., and Parkinson, B., "Aircraft Applications of GPS-Based Attitude Determinations," pp. 775–782; also *Proceedings of the ION GPS—92* (Albuquerque, NM), Institute of Navigation, Washington, DC, Sept. 16–19, 1992.

[11] Anon., "Minimum Aviation System Performance Standards DGNSS Instrument Approach System: Special Category I (SCAT I)," Radio Technical Commission for Aeronautics, RTCA DO-217 Washington DC, Aug. 27, 1993.

Chapter 14

Aircraft Automatic Approach and Landing Using GPS

Bradford W. Parkinson* and Michael L. O'Connor†
Stanford University, Stanford, California 94305
and
Kevin T. Fitzgibbon‡
São Jose Dos Campos, Brazil

I. Introduction

A. Autolanding Conventionally and with GPS

MOST conventional aircraft automatic landing systems use the instrument or microwave landing systems (ILS or MLS) in their terminal approach phases. These systems supply the autopilot with the aircraft's *angular* deviation from a desired flight path, which essentially corresponds to the measurement of vertical and lateral positions. Basic ILS can only satisfy the FAA's nonprecision and Category I landing requirements; and aircraft using the MLS or an improved ILS system can land with Category III required accuracy. In these autopilots, velocity is typically calculated by filtering and differentiating position, or by integrating the acceleration outputs of an Inertial Measurement Unit. Some disadvantages of these integrated systems include their high user costs and dependence on expensive ground equipment. Also, the noisy ILS and MLS signals are typically processed with a smoothing filter. This causes lags that must be compensated for by the landing autopilot.

Unlike most other navigation aids, a *GPS receiver directly measures three-dimensional velocity* with extreme accuracy (for DGPS and CDGPS,§ better than 5 cm/s). For the autopilot designer, this is of great value. A *direct* measurement of true ground speed not only assists in normal landings, it gives important advanced knowledge of wind gusts and shears. In addition to velocity, a single, state-of-the-art, GPS receiver can provide accurate three-dimensional position

*Professor, Department of Aeronautics and Astronautics; Director of the GPS Program.

†Research Assistant, Department of Aeronautics and Astronautics.

‡Consultant and Professor of Aeronautics.

§DGPS is differential GPS, which is covered in Chapter 1 of this volume. CDGPS is carrier-phase differential GPS.

and attitude measurements. *The power of GPS is that a single electronic box can measure three-dimensional position, velocity, and attitude for all phases of flight, including precision landing.* Traditionally, full autolanding requires an expensive inertial navigator. The results presented herein suggest that a simple system of rate gyros may be all that is required for a Category II or possibly a Category III landing. Of course, further effort is required to ensure that integrity specifications are met, which would include determining the level of redundancy and reliability required.

B. Simulations Results Presented

In this chapter, an automatic landing system is designated and simulated for a Boeing 747 using a discrete-time controller and an optimal estimator, which both rely on GPS sensors. The technique for designing the autopilot is fully described; the appendices include the parameters used. Four alternative sets of sensors are included in the simulations:

1) standard GPS;
2) standard GPS augmented with a radar altimeter;
3) code differential GPS (DGPS) without radar altimeter; and
4) carrier-phase differential GPS (CDGPS).

A block diagram of the landing system is presented in Fig. 1, where the radar altimeter and differential aids are optional.

The autopilot controller is implemented in two ways. The first is a fairly standard linear quadratic Gaussian (LQG) regulator, which estimates wind disturbances directly. The second uses an integral control law (ICL), which does not directly estimate the winds, but includes integral states that "soak up" the output errors. Optimal estimator theory, LQG regulator theory, and representative aircraft models are presented in many textbooks.[1-4] Holley and Bryson[5] presented a modified integral control design for multi-input multi-output *continuous* autolanding systems. They applied their results to an aircraft lateral mode in the presence of constant crosswinds.

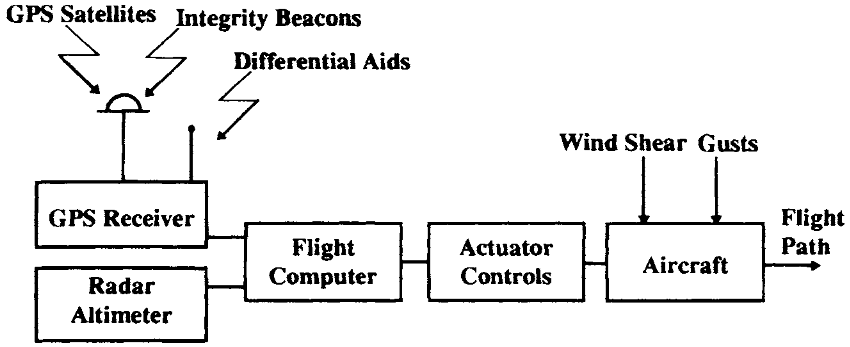


Fig. 1 The GPS autoland system block diagram.

The results of the simulations presented here are for the *discrete* design version of this ICL, and are applied to a *complete* aircraft model with both lateral and longitudinal modes subjected to wind shear and gust disturbances. Disturbance models for the simulation and their numerical values were extracted from Holley and Bryson,[5] Roskam[6] and Bryson.[2]

The simulation results show that under normal wind conditions and typical satellite geometries, GPS and DGPS can easily meet the FAA navigation system accuracy required for a nonprecision approach. Augmenting stand-alone GPS with a radar altimeter meets the accuracy requirements for a precision Category I approach. Utilizing carrier-phase measurements with CDGPS meets the required navigation system accuracy for a precision Category III approach *without an inertial navigation system.* Comparison of the two control laws suggests that the ICL controller is more robust to unexpected variations in disturbance inputs than the LQG controller.

II. Landing Approach Procedures

A. Instrument and Microwave Landing Systems

The most common landing system currently in use (1994) is ILS. MLS is the most recently developed system with improved accuracy over ILS. Both systems provide a reference path to the aircraft in terms of an azimuth and an elevation. Because of the nature of these angular measurements, position errors with these systems vary with distance from the ground-based transmitters. Very often these systems are aided by onboard inertial navigation systems that provide additional information about attitude, position, and velocity. The aircraft is controlled in order to keep its path within a reference cone that guarantees the position and velocity accuracies required at touchdown. Table 1 shows FAA landing requirements at the time of this writing. GPS is providing a continuum of accuracy that has led to the development of new types of specifications called the *tunnel concept.*

Figure 2 presents a sketch of the typical flight phases for an ILS landing. These are: 1) the initial approach; 2) the glide-slope phase; and 3) the flare phase. This nomenclature belongs to the ILS/MLS systems and is preserved and used in this work.

Table 1 FAA navigation system accuracy standards

Operational phase	Minimum altitude	Accuracy lateral, 2 drms	Vertical, rms
En route terminal	152 m	7400 m	500 m
Approach landing			
Nonprecision	76.2 m	3700 m	100 m
Precision Category I	30.5 m	9.1 m	3.0 m
Precision Category II	15.2 m	4.6 m	1.4 m
Precision Category III	0 m	4.1 m	0.5 m

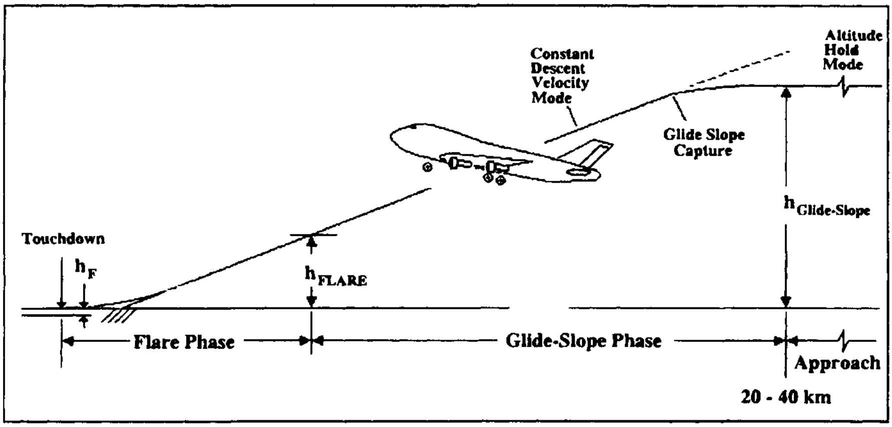


Fig. 2 Typical longitudinal approach path.

1. Initial Approach

During the *initial approach* phase, the aircraft starts at cruising altitude and descends to a lower altitude between 500 and 1500 m, which occurs at a distance of less than 40 km from the runway. After this descent, the aircraft enters an *altitude hold* mode. From this condition, the aircraft is able to capture the ILS/MLS radio signals and follow an accurate path toward the runway. Existing navigation equipment and autopilots can bring an aircraft within 150 m in position and less than 10 degs in azimuth accuracy with respect to the runway's position and azimuth during initial approach, even in the presence of disturbing winds.

2. Glide Slope

When the initial approach path intersects the desired glide path, the aircraft enters a constant-descent or *glide-slope* mode in which the altitude rate is kept between -2 and -3 m/s. This typically results in a path inclination between 2 and 3 degs. The transition maneuvers are designed to be safe and comfortable to passengers with accelerations not exceeding 0.15 g. During the glide slope, the autopilot keeps the aircraft deviation from the center of the ILS/MLS radio beam as small as possible.

3. Flare

The *flare phase* starts at a switching altitude h_{FLARE}, which depends upon the glide-slope altitude rate and the desired altitude rate for aircraft touchdown on the runway This last phase ends with touchdown where the altitude rate should be about -0.5 m/s. The autopilot performs this maneuver by flying an asymptotic approach toward a final altitude (h_F) chosen to be slightly below the runway. An approach to the exact runway altitude is not desirable, because it would greatly magnify small positioning errors at touchdown. Also, we see in the simulation that a sensor bias greater than h_F can cause the aircraft never to reach the runway.

B. GPS Approach

The GPS system is an independent position, velocity, and attitude sensor with no ground aiding equipment in the sense that no reference beam is provided by the ground equipment to the autopilot. Unlike ILS and MLS, *any* convenient reference path can be created with a GPS system based upon the runway's known position. Integrity beacons are used with CDGPS to calibrate satellite integer ambiguities and provide an additional carrier-phase reference. These simple, inexpensive devices also provide integrity: calculations show that they can meet the FAA requirement of less than one misleading position in a billion landings. Although the aircraft is required to fly in the vicinity of these transmitters to reliably achieve centimeter-level accuracies, they do *not* restrict aircraft motion to a particular reference path for landing.

In the examples presented, we have used a path similar to the ILS and MLS systems for landing with GPS. Other approach paths could be used. For example, a parabolic, continuous-are descent, combining phases 2 and 3 of the conventional approach, may afford advantages and would be very easy for the GPS-equipped aircraft to fly.

III. Aircraft Dynamics and Linear Model

In the simulation, the aircraft is modeled with six degrees of freedom in small perturbations around a stable equilibrium point. The particular steady-state equilibrium point is the landing configuration at sea level for a Boeing 747 as derived from Bryson.[2] The components of the state vector are the aircraft position, velocity, attitude, attitude rate, and thrust specific force. The components of the control vector are the elevator, aileron, and rudder deflections, as well as the thrust specific force command. The disturbances are the longitudinal, vertical, and lateral winds.

The aircraft controller is assumed to have a perfect model of plant dynamics and sensor characteristics. However, the wind model for the simulation is inexact for both the LQG and ICL control law designs. The differences are described below.

A. State Vector

The simulations are performed in the state-space domain and the components of the state vector $\boldsymbol{X(t)}$ are as follows:

$$\begin{aligned}\boldsymbol{X(t)}^T = [&u(t)\ w(t)\ q(t)\ \theta(t)\ d(t)\ h(t)\ x(t)\ \delta_T(t)\ U_0(t)\\ &|\ v(t)\ r(t)\ p(t)\ \phi(t)\ \psi(t)\ y(t)]\end{aligned} \tag{1}$$

where,

Longitudinal mode:

u = longitudinal groundspeed
w = vertical groundspeed
q = pitch rate
θ = pitch attitude angle

d = vertical deviation from glide-slope
h = altitude
x = longitudinal displacement
δ_T = thrust specific force
U_o = nominal forward speed (doesn't change—used in calculation of d)

Lateral mode:

v = lateral groundspeed
r = yaw rate
p = roll rate
ϕ = roll attitude angle
Ψ = yaw attitude angle
y = lateral displacement

B. Control Vector

The components of the control vector $\boldsymbol{U(t)}$ are as follows:

$$\boldsymbol{U(t)}^T = [\delta_E(t)\ \delta_{TC}(t) \mid \delta_A(t)\ \delta_R(t)] \tag{2}$$

where,

Longitudinal mode:

δ_E = elevator deflection
δ_{TC} = commanded thrust specific force

Lateral mode:

δ_A = aileron deflection
δ_R = rudder deflection

C. Disturbance Vector

The components of the disturbance (wind) vector $\boldsymbol{W(t)}$ are as follows:

$$\boldsymbol{W(t)}^T = [W_U(t)\ W_W(t) \mid W_V(t)] \tag{3}$$

where,

Longitudinal mode:

W_U = longitudinal wind
W_W = vertical wind

Lateral mode:

W_V = lateral wind

D. Measurement Vector

During the glide-slope phase, the components of the measurement vector $\boldsymbol{Z(t)}$ are as follows:

$$\boldsymbol{Z(t)}^T = [u(t)\ w(t)\ \theta(t)\ d(t) \mid v(t)\ \phi(t)\ \psi(t)\ y(t)] \tag{4}$$

E. Equations of Motion

The first-order differential equations of motion of the aircraft model are put into the continuous state-space representation,

$$\mathrm{d}X(t)/\mathrm{d}t = AX(t) + BU(t) + B_W W(t) \tag{5}$$

with the following output and measurement equations,

$$Y(t) = CX(t) \tag{6}$$

$$Z(t) = HX(t) + \mu(t) \tag{7}$$

$W(t)$ and $\mu(t)$ are assumed to be uncorrelated white Gaussian noises with given means and variances in the ICL controller design. In the LQG controller, the estimated wind vector ***We(t)*** uses an exponentially correlated model represented by the following state-space equation:

$$\mathrm{d}\boldsymbol{We(t)}/\mathrm{d}t = -1/\tau_W\ \boldsymbol{We(t)} + r(t) \tag{8}$$

The time constant τ_w chosen is large compared to the time constants of the aircraft, so the wind estimate is very close to the actual wind bias. $\boldsymbol{We(t)}$ is appended to the estimated state vector $\boldsymbol{Xe(t)}$ for estimator design, creating the following augmented state vector:

$$Xe_A(t) = \begin{bmatrix} \boldsymbol{Xe(t)} \\ \boldsymbol{We(t)} \end{bmatrix} \tag{9}$$

F. Wind Model

The true wind disturbances are modeled in three dimensions with respect to the airframe of the aircraft. In each direction, the wind is composed of a random but correlated gust component and a steady shear component. Table 2 presents the correlation times used for each wind gust component. For the nonsteady wind disturbances, the following exponentially correlated model is used:

$$\mathrm{d}W(t)/\mathrm{d}t = -1/\tau\ W(t) + v(t) \tag{10}$$

Table 2 Wind gust disturbance correlation times

Disturbance	$1/\tau$ (s^{-1})
Longitudinal wind	0.43
Vertical wind	1.06
Lateral wind	0.14

where

W = W_U, W_v or W_W
v = white Gaussian noise
τ = true correlation time

The continuous disturbance covariance matrix $W_C = E\{v^2\}$ is chosen so that after discretization, the wind covariance $E\{W^2\} = 0.7$ m/s in all three axes.[2]

Steady winds in the simulation are modeled as a function of altitude. The wind intensity in each direction varies linearly from altitude zero (runway's altitude) up to a steady constant value. Two parameters characterize each wind profile:

1) intensity W_{SAT} (constant saturation value); and
2) intensity gradient D_H

or mathematically, as follows:

$$W_{\text{BIAS}}(t, h) = D_H h(t), \qquad h(t) \leq W_{\text{SAT}}/D_H \tag{11}$$

$$W_{\text{BIAS}}(t, h) = W_{\text{SAT}}, \qquad h(t) > W_{\text{SAT}}/D_H \tag{12}$$

G. Throttle Control Lag

The following equation was used to model the lag in throttle control as a first-order process with a 4-s time constant. This has an appreciable effect, especially during the flare phase of landing.

$$\mathrm{d}\delta_T/\mathrm{d}t = -0.25(\delta_T - \delta_{TC}) \tag{13}$$

H. Glide-Slope Deviation

Typically, the equations of motion for an aircraft are given for a vehicle in straight and level flight. However, during the glide-slope phase we are interested in controlling the deviation from a nominal trajectory that is not horizontal. Implementing a controller that follows a ramp input in altitude h is one solution, but it turns out to be simpler to define a new variable d, which is the perpendicular distance from the glide slope. The equations derived in Bryson's text[2] give a differential equation for h:

$$\mathrm{d}(h)/\mathrm{d}t = -w \cos\theta + (U_0 + u)\sin\theta \tag{14}$$

or

$$\mathrm{d}(h)/\mathrm{d}t \simeq -w + U_0\theta \tag{15}$$

Using this, we can take the glidè-slope angle γ into account and compute the differential equation for d:

$$\begin{aligned}\mathrm{d}(d)/\mathrm{d}t = {} & [-w\cos\theta + (U_0 + u)\sin\theta]\cos\gamma \\ & + [-w\sin\theta + (U_0 + u)\cos\theta]\sin\gamma\end{aligned} \tag{16}$$

or

$$\mathrm{d}(d)/\mathrm{d}t \simeq -w\cos\gamma + U_0\theta\cos\gamma - u\sin\gamma + U_0\sin\gamma \tag{17}$$

In the simulation truth model, both h and d are computed. The controller operates on d in the glide-slope phase and h in the flare phase. This causes some difficulty in controller design. For example, care must be taken in transition from glide-slope to flare phase, because the state variable changes (see Appendix A). Also, the new state variable (U_0), which does not change with time, must be added to the dynamic equations to account for the constant fourth term of Eq. (17).

IV. Autopilot Controller

A. Linear Quadratic Gaussian and Integral Control Law Controllers

Two different types of autopilot controllers were used in these simulations. The first is based on a standard LQG regulator that has extra states for direct estimates of wind disturbances. The second is based on an ICL regulator that does not directly estimate the winds but includes added states for the integral of the output errors.

During the glide-slope phase, the controller works to correct the perpendicular deviations from the desired flight path. During flare, it tries to keep the aircraft altitude on the exponential path described above while driving the lateral displacement to zero. For this reason, both controllers were actually designed with the capability of adjusting to nonzero set points. A detailed description of how this was done can be found in Appendix A.

Each of the two methods of controller design has its advantages and limitations. The LQG regulator is truly optimal in minimizing the quadratic cost function (described below) when the plant, control, and disturbance models are known exactly. If the controller model differs significantly from the truth model, rms performance is degraded, and nonzero disturbances and set-points can lead to a nonzero steady output offset. The ICL controller slightly degrades the closed-loop system performance compared to the LQG controller when the controller model matches the truth model. However, the ICL control law has the ability to compensate for nonzero steady disturbances of (possibly) unknown origin, so modeling errors do not result in a steady output error.

B. Regulator Synthesis

During the landing phases, the aircraft's attitude, position, velocity, and controls are limited either because of physical constraints, such as the maximum available rudder deflection, or constraints that ensure structural safety and passenger comfort. The latter constraints usually include accelerations and attitude angle limits. Table 3 presents the typical maximum values for a Boeing 747 aircraft. These maximum values affect the optimal controller gains for the glide-slope and flare phases of flight.

The continuous system presented in Eqs. (5–7) can be discretized and represented by the following state, output, and measurement equations:

$$X_{K+1} = A_D X_K + B_D U_K + B_{WD} W_K \tag{18}$$

$$Y_K = C X_K \tag{19}$$

$$Z_K = H X_K + \mu_K \tag{20}$$

Table 3 Maximum limits for states and controls for the B747 on landing

Variable	Units	Glide slope	Flare
Pitch rate	deg/s	*	*
Pitch attitude	deg	*	5
Roll rate	deg/s	5	5
Roll attitude	deg	15	5
Yaw rate	deg/s	*	*
Yaw attitude	deg	10	5
Lat. displacement	m	15	8
Lat. velocity	m/s	*	*
Long. displacement	m	*	*
Long. velocity	m/s	*	*
Vert. displacement	m	8	1.5
Sink rate	m/s	3	0.6
Elev. deflection	deg	5	5
Aileron deflection	deg	5	5
Rudder deflection	deg	5	5
Throttle specific force	m/s^2	1	1

Constraints were not placed on the asterisk quantities.

where the time index K refers to the time $t = KT$ (T = sampling period), and A_D, B_D, B_{WD}, C, and H are the discrete transition, control, disturbance, output, and measurement distribution matrices, respectively. The optimal regulator is designed to minimize the following cost function:

$$J = \Sigma(X_K^T Q_D X_K + 2 X_K^T N_D U_K + U_K^T R_D U_K) \tag{21}$$

where Q_D, R_D, and N_D are the discrete weighting matrices of the states, controls, and their correlated terms. The discrete weighting matrices can be obtained by discretizing the continuous weighting matrices Q and R. N_D is caused by the coupling resulting from the discretization and is generally nonzero, even if there is no coupling in the continuous case. The following relations are used:

$$Q_D = \int_0^T [F^T(t)QF(t)]\mathrm{d}t \tag{22}$$

$$R_D = RT + \int_0^T [G^T(t)QG(t)]\mathrm{d}t \tag{23}$$

$$N_D = \int_0^T [F^T(t)QG(t)]\mathrm{d}t \tag{24}$$

where

$$F(t) = \exp(At)$$

$$G(t) = \int_0^t F(t - \tau)B\mathrm{d}\tau$$

The Q and R matrices can be defined as diagonal matrices, and the diagonal elements are defined using the following rule-of-the-thumb method (sometimes called Bryson's rule):

$$Q = \mathrm{diag}[1/(X_{1\max})^2 \cdots 1/(X_{N\max})^2] \tag{25}$$

and

$$R = \mathrm{diag}[1/(U_{1\max})^2 \cdots 1/(U_{M\max})^2] \tag{26}$$

where $X_{i\max}$ and $U_{j\max}$ are the maximum values that each variable is allowed to reach in the dynamic or steady environment. These values are usually determined by such physical limitations as available power, limited angular deflections, and safety and structural failure requirements. The following steady-state control law results:

$$U_K = -C_X X_K \tag{27}$$

V. GPS Measurements

A key factor in the design of the autolanding system is a realistic set of GPS biases and noise. These errors can vary considerably among the various equipment manufacturers. The values used to evaluate these quantities are the current (1994) state-of-the-art in accuracy. The rms measurement error and biases of typical GPS [without selective availability (SA)], GPS aided by a radar altimeter, DGPS, and carrier-phase DGPS systems are presented in Table 4. GPS was also used for attitude determination in the simulation, with a 0 deg bias error and 0.2 deg standard deviation.

Figure 3 details the landing system as implemented with the integral control law, including the GPS receiver, the estimator, and the digital controller. The measurements provided by the receiver are sampled at discrete time intervals and sent to the estimator. The function of the estimator is to combine the limited sensor information with the known plant model to generate estimates of all the

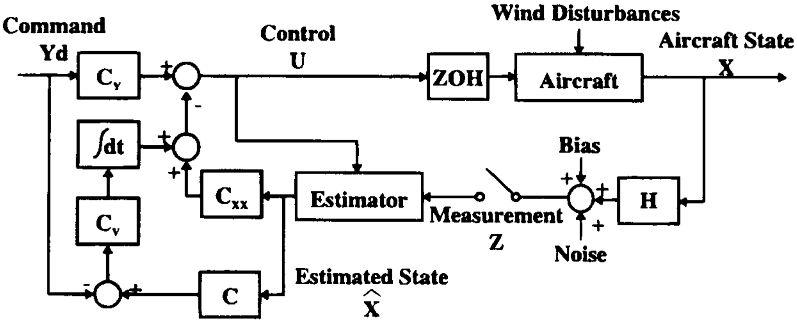


Fig. 3 Integral control law simulated block diagram.

Table 4 1σ measurement error and biases of typical GPS and DGPS

	GDOP[a]		UERE,[d] m		UERRE[e] m/s		Horizontal pos. error, m		Vertical pos. error, m		Horizontal vel. error, m/s		Vertical vel. error, m/s	
System	HDOP[b]	VDOP[c]	Bias	σ	Bias	σ	Bias	σ	Bias	σ	Bias	σ	Bias	σ
GPS (no SA)	3.0	5.0	3.0	0.5	0.02	0.01	9.0	1.5	15.0	2.5	0.06	0.03	0.10	0.05
GPS + Alt.	3.0	N/A	3.0	0.5	0.02	0.01	9.0	1.5	0.2	0.1	0.06	0.03	0.05	0.025
DGPS	3.0	5.0	1.5	0.5	0.01	0.005	4.5	1.5	7.5	2.5	0.03	0.015	0.05	0.025
CDGPS	3.0	5.0	0.01	0.002	0.005	0.005	0.03	0.006	0.05	0.010	0.015	0.015	0.025	0.025

[a] GDOP = geometric dilution of precision.

[b] HDOP = horizontal dilution of precision.

[c] VDOP = vertical dilution of precision.

[d] UERE = user equivalent range error.

[e] UERRE = user equivalent range rate error.

state variables (see Appendix B). The controller calculates the control commands based on the estimated states using full state feedback. The control signals are sent to a zero-order-hold (ZOH) digital-to-analog converter and into the aircraft as a continuous signal.

VI. Results

A. Cases Simulated

The following cases were simulated:

1) standard GPS, GPS with altimeter, DGPS, and CDGPS;

2) linear quadratic Gaussian and integral control law regulators; and

3) Glide-slope and flare phases.

The numerical results of the simulation are presented in tables with the statistical mean and standard deviation of the flight path errors. Plots of the altitude, lateral displacements, and altitude rates during a typical CDGPS landing are also presented.

Tables 5–10 give the statistics of the lateral and vertical errors for GPS alone, GPS with altimeter, DGPS, and CDGPS configurations. Tables 5 and 6 show the total system error during the approach, while Tables 7–10 break this error down into the *navigation system error* (the difference between actual and estimated position), and the *flight technical error* (the difference between estimated and desired position). The altitude rates at touchdown and landing success rates are also presented in the Tables 5 and 6. The landing success rate is important because in some simulation runs, the stand-alone GPS sensor bias was large enough to keep the aircraft from ever touching down.

To create the tables, 60 landings were simulated for each configuration with a glide-slope angle of 2.5 deg, $h_{\text{FLARE}} = 15$ meters, $h_F = 1.2$ m, and $1/a = 5.7$ s (see equation A10). The measurement sample frequency was 10 hertz. Sensor bias errors were recalculated along with random errors for each run. The tables show the mean and standard deviation values for all runs combined.

Glide-slope acquisition was performed at the low altitude of 200 m to keep a reasonably short simulation length. There was a significant transient at the

Table 5 Total system error for linear quadratic Gaussian controller

	GPS alone (no SA)		GPS + altimeter		DGPS		CDGPS	
Displacement	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Glide-slope phase								
Vertical, m	4.90	14.76	0.40	1.54	2.11	7.58	0.30	0.95
Lateral, m	2.86	8.56	2.81	8.46	1.55	4.50	0.37	1.62
Flare phase								
Vertical, m	9.58	11.33	1.58	1.24	4.89	6.17	0.50	0.75
Lateral, m	1.12	8.87	0.09	8.06	0.01	4.64	−1.85	1.13
Alt. rate, m/s	−1.81	1.64	−0.91	0.43	−1.33	1.00	−0.71	0.34
% Successful landings	48		100		52		100	

Table 6 **Total system error for integral control law controller**

	GPS alone (no SA)		GPS + altimeter		DGPS		CDGPS	
Displacement	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Glide-slope phase								
Vertical, m	0.78	14.54	−0.25	0.89	1.19	7.49	0.29	0.73
Lateral, m	3.00	8.38	2.83	8.29	1.20	4.35	−0.13	1.39
Flare phase								
Vertical, m	6.66	11.39	−0.14	0.72	3.68	6.25	−0.11	0.58
Lateral, m	4.44	9.02	3.40	8.40	1.74	4.78	−0.28	1.62
Alt Rate, m/s	−2.50	1.95	−0.61	0.30	−1.19	1.09	−0.56	0.27
Successful landings	57		100		63		100	

beginning of each simulation while the wind bias compensation built up. In the LQG case, the initial estimates of wind biases were zero, whereas, in the ICL case, the initial integral error terms were zero. Tables 5–10 show the results of data taken after the effects of this initial transient have settled out.

B. Landing with GPS Alone

For autoland using nondifferential GPS, there is little difference in performance between the LQG and ICL controllers. Tables 5 and 6 show that the vertical rms position errors lie between 10 and 15 m, while lateral rms position errors are around 8 or 9 m. Tables 7 and 8 show that these errors are primarily caused by navigation system error, not flight technical error. In other words, the imprecise GPS measurements are the main cause of the total system errors, not the autopilot controller. From these results, we see that the GPS system in the absence of SA clearly *meets the FAA nonprecision approach requirements* shown in Table 1, but the vertical position errors exceed the Category I precision landing specification.

C. Landing with GPS plus Altimeter

Specifications for precision landing approach are more stringent in the vertical dimension than in the horizontal dimension. Unfortunately, GPS horizontal measurements are typically more accurate than vertical measurements because of satellite geometry. One solution to improve vertical navigation accuracy is to augment GPS with a radar altimeter. Although this sensor does nothing to improve lateral accuracies, Tables 5 and 6 show that vertical rms position errors are reduced to around 1 m. Once again, the total system error is dominated by the navigation system error for both the LQG and ICL controllers. This combined sensor system *meets the accuracy specifications for the FAA precision Category I approach,* with extremely good vertical navigation system error compared to GPS alone.

D. Landing with Differential GPS

Another method for improving the accuracy of the stand-alone GPS signal is to use code-differential corrections from a nearby reference station. This has the advantage of improving both lateral and vertical measurements of position and velocity. Tables 7 and 8 show that DGPS offers a significant improvement in the navigation system accuracy of both autopilots, as expected. It is interesting to note that the addition of reference corrections also improves the flight technical error, because wind gusts and biases are now better estimated (LQG) or otherwise accounted for (ICL). Tables 7 and 8 show that lateral and vertical position estimate errors during flare were reduced from 12 or 13 m with GPS to around 7 m with DGPS. The results suggest that under the wind conditions described for this simulation, *DGPS alone meets the FAA Category I precision landing lateral error requirement, but exceeds the vertical requirement of 3 m.* The vertical navigation error bias was responsible for many unsuccessful landings, because the flare asymptote was only 1.2 m below the runway in these simulations.

E. Landing with Carrier-Phase

A third method for improving sensor accuracy during aircraft landing is to perform real-time carrier-phase differential GPS. As shown in Table 4, this method

Table 7 Navigation system error for linear quadratic Gaussian controller

	GPS alone		GPS + altimiter		DGPS		CDGPS	
Displacement	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Glide-slope phase								
Vertical, m	4.51	14.76	0.08	0.36	1.80	7.52	0.02	0.08
Lateral, m	2.31	8.22	2.23	8.13	1.14	4.10	0.01	0.03
Flare phase								
Vertical, m	9.05	12.57	0.33	0.32	4.49	6.74	0.03	0.07
Lateral, m	2.44	8.67	2.37	7.96	1.34	4.42	0.00	0.03

Table 8 Navigation system error for integral control law controller

	GPS alone		GPS + altimeter		DGPS		CDGPS	
Displacement	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Glide-slope phase								
Vertical, m	0.30	14.59	−0.69	0.31	0.78	7.49	−0.11	0.06
Lateral, m	3.19	8.21	3.01	8.12	1.37	4.10	0.05	0.03
Flare phase								
Vertical, m	7.34	12.77	−0.07	0.31	4.08	6.86	−0.01	0.05
Lateral, m	4.14	8.70	3.66	8.12	1.72	4.44	0.02	0.03

offers the tremendous advantage of raw measurement errors that are much smaller than stand-alone GPS. The total position error for an aircraft landing with GPS, GPS with altimeter, or DGPS was primarily dominated by navigation system error. *The simulation shows that the system error of a Boeing 747 landing with CDGPS is dominated by the aircraft flight technical error;* i.e. the ability of the aircraft to follow a known trajectory in the presence of external physical disturbances. This means the sensor measurements of position, velocity, and attitude are so accurate that autopilot performance is basically determined by actuator control authority and passenger safety and comfort. The navigation system error for CDGPS shown in Tables 9 and 10 *easily meets the FAA precision Category III accuracy requirements.*

Figure 4 shows a typical landing using CDGPS in the presence of wind gusts and steady wind disturbances. Both altitude plots show the transient at the beginning of the simulation where the wind bias compensation is building up. The lateral displacement plots show the performance of each controller in the presence of wind gusts alone and in the presence of wind gusts and a wind bias. The initial 7.5 m/s "step" in wind bias leads to a transient with about a 5 m maximum lateral error. All four simulations were run with the same initial conditions and the same random errors for sensors and disturbances.

Figure 5 shows the altitude and altitude rate for typical CDGPS landings using the LQG and ICL control laws. A landing with no winds is compared to a landing

Table 9 Flight technical error for linear quadratic Gaussian controller

	GPS alone		GPS + altimeter		DGPS		CDGPS	
Displacement	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Glide-slope phase								
Vertical, m	0.39	1.51	0.32	1.28	0.32	1.12	0.28	0.93
Lateral, m	0.55	2.10	0.57	2.13	0.41	1.75	0.36	1.62
Flare phase								
Vertical, m	0.54	2.07	1.25	1.13	0.40	1.24	0.47	0.76
Lateral, m	−1.32	1.93	−2.27	1.57	−1.33	1.51	−1.85	1.13

Table 10 Flight technical error for integral control law controller

	GPS alone		GPS + altimeter		DGPS		CDGPS	
Displacement	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Glide-slope phase								
Vertical, m	0.47	0.79	0.44	0.75	0.41	0.72	0.39	0.71
Lateral, m	−0.19	1.59	−0.18	1.58	−0.17	1.40	−0.18	1.38
Flare phase								
Vertical, m	−0.68	1.74	−0.07	0.59	−0.40	0.97	−0.10	0.57
Lateral, m	0.30	2.22	−0.27	1.85	0.02	1.78	−0.30	1.62

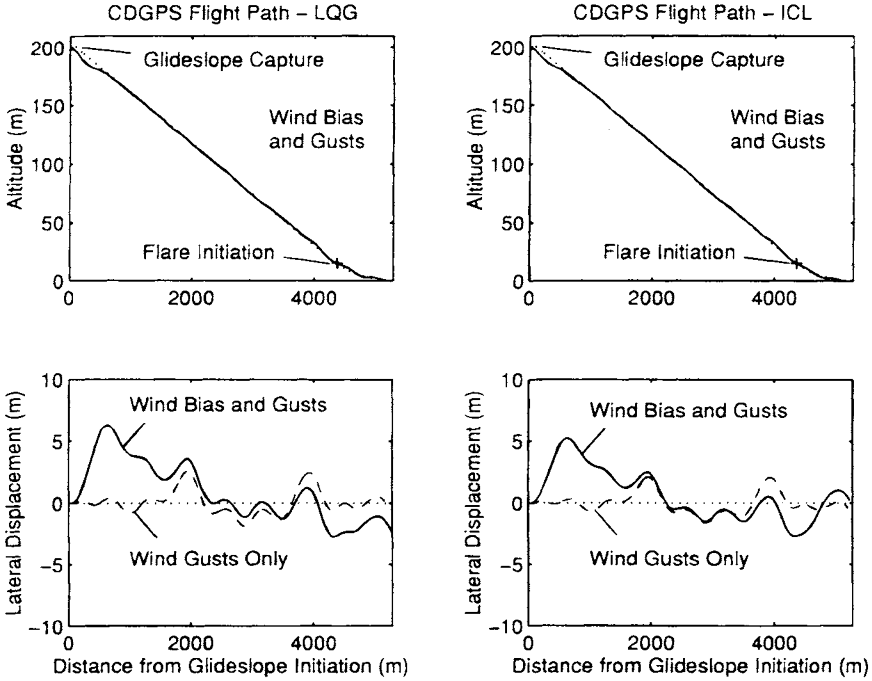


Fig. 4 Typical CDGPS approach.

with both wind biases and gusts. The goal to land with an altitude rate of −0.5 m/s is easily met with no winds present. As seen in Tables 5 and 6, this goal is slightly exceeded in the presence of winds. The plots of Fig. 5 suggest that errors in altitude rate are primarily caused by the wind gusts rather than the wind bias, which has fallen off appreciably at this low altitude.

F. Linear Quadratic Gaussian vs Integral Control Law

As expected, *the navigation system errors for both controller types are very similar.* The primary difference between the LQG and ICL controllers is seen in the flight technical error.

From the simulation results shown in Tables 5 and 6, we can see that the LQG and ICL controllers have approximately the same total system error *standard deviations.* For example, using DGPS, Table 5 shows the LQG vertical accuracy standard deviation is 7.58 m during glide slope and 6.17 m during flare. These are similar to the ICL values in Table 6, which are 7.49 m and 6.25 m, respectively. Although the LQG controller was the "optimal" design for the given cost matrices, the ICL controller was able to achieve comparable performance *in response to random zero-mean inputs.*

In fact, the ICL controller actually performed better than the LQG controller in response to *nonzero wind biases.* From Tables 9 and 10 we see that the LQG *mean* flight technical error is relatively high, especially during flare. For example,

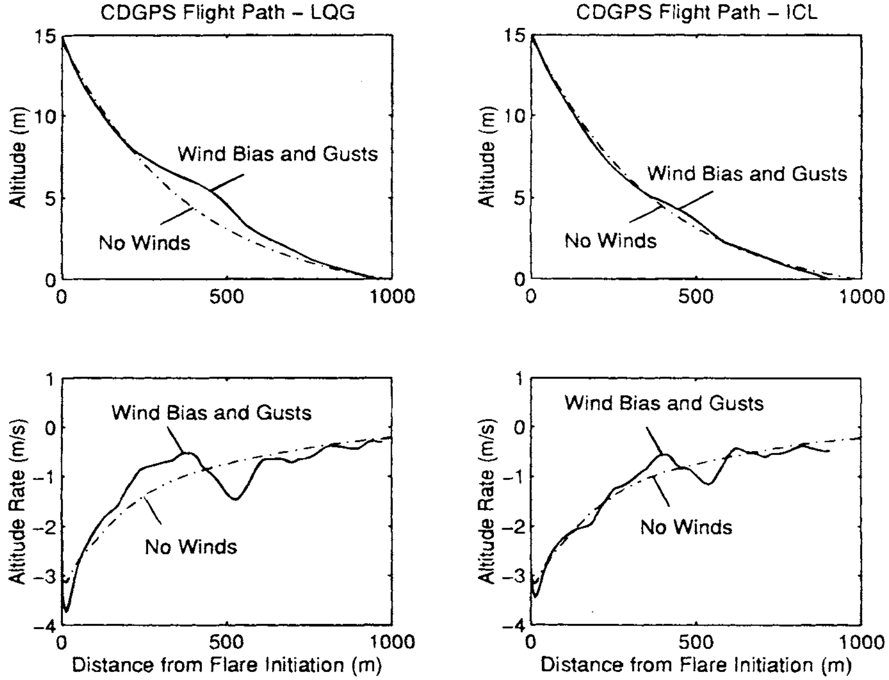


Fig. 5 Typical CDGPS landing.

using CDGPS during flare, the LQG bias errors are 0.47 m vertically and −1.85 m laterally. The ICL controller results in mean errors of −0.10 m vertically and −0.30 m laterally. During flare, the wind bias is varying linearly with altitude. This suggests that the ICL control law is better adapted than the LQG controller to compensate for a changing, nonzero wind bias. Figure 4 also suggests that the LQG controller is more susceptible to variations in wind bias disturbances, because the initial lateral position transient is larger than for the ICL controller.

VII. Conclusions and Comments

Autopilot design based on GPS has several advantages over ILS- and MLS-based systems. GPS is clearly less expensive because it does not rely on costly equipment on the ground or in the air. It is also more flexible, because approaches are not confined to take place within a narrow radar beam. *Most importantly, the GPS sensor measurements are fundamentally better suited for use by an autopilot.* ILS and MLS measurements are based on angular deviation from a desired flight path. They, therefore, have changing sensitivity to position errors as the aircraft comes closer to the transmitter. The GPS measures position in three dimensions with no real degradation in accuracy nor any change in sensitivity through touchdown. GPS also offers the enormous advantage of highly accurate velocity measurements. The estimator can use this information to improve its position estimates and to better determine the magnitude of wind disturbances. This is particularly helpful in the event of wind shears near the ground.

As with any computer simulation, there are some limitations to the study presented here. For example, the simulation assumes approximately average wind conditions and typical GPS sensor errors. More severe winds or poorer GPS performance (such as a higher GDOP or a less accurate receiver) would produce worse results, although the conclusions about CDGPS would probably still be valid. Also, we have assumed that the aircraft dynamics are linear in the region of operation with no significant cross coupling between the lateral and longitudinal modes. Finally, we have assumed that the computer has perfect knowledge of the linear plant model and sensor characteristics and that corrections from the ground reference station when applicable, are continuously available. All of these assumptions are reasonable for the initial autopilot design. To ensure robustness, further analysis, backed by flight tests must be undertaken.

Simulation results suggest the following conclusions based upon the previous assumptions:

1) The ICL control law seems to be more robust to variations in design parameters than a standard LQG controller. In particular, the ICL is less sensitive to variations in nonzero wind disturbances.

2) Nondifferential GPS, even in the absence of SA, can only satisfy the FAA nonprecision landing requirement.

3) Nondifferential GPS augmented with a radar altimeter meets the FAA Category I precision landing requirement.

4) Differential GPS (code phase) meets the lateral accuracy but exceeds the vertical accuracy requirement for FAA Category I precision landing.

5) Carrier-phase differential GPS meets the accuracy specification for FAA Category III precision landing. Unlike the other three cases, position errors for CDGPS are dominated by autopilot error, not sensor uncertainty.

The only four sensor configurations examined in this study were GPS alone, GPS with radar altimeter, DGPS alone, and CDGPS alone. In all cases, system accuracy (and integrity) could be improved by using additional sensors. A simple inertial measurement unit would provide redundancy and improve estimates of aircraft position and heading. Also, because GPS is a highly accurate sensor of *ground* speed, the addition of an *air* speed sensor could greatly improve the estimates of wind speed—especially wind shear.

Appendix A: Discrete Controllers

Linear Quadratic Gaussian Controller for Constant Set Points and Steady Wind Disturbances

If we want the output vector Y_K to take on specific values (Y_D), we can often look at the linear state equations as perturbations about a nonzero steady value. When doing this, we must be careful that the linearized equations still hold in the new regime.

The steady-state values are defined as follows:

$$X_{SS} = X_{SS}(K) = X_{SS}(K + 1) \tag{A1}$$

$$W_{SS} = W_{SS}(K) = W_{SS}(K + 1) \tag{A2}$$

$$Y_{SS} = Y_D \tag{A3}$$

Substitute into Eqs. (18) and (19) to obtain the following:

$$-B_{WD} W_{SS} = [A_D - I]X_{SS} + [B_D]U_{SS} \tag{A4}$$

$$Y_D = [C]X_{SS} + [0]U_{SS} \tag{A5}$$

where I is the identity matrix. For arbitrary nonzero W_{SS} and Y_D, the system has a solution if the rank of M, a matrix whose elements are defined by,

$$\begin{bmatrix} X_{SS} \\ U_{SS} \end{bmatrix} = \begin{bmatrix} M_{XW} & M_{XY} \\ M_{UW} & M_{UY} \end{bmatrix} \begin{bmatrix} -B_{WD}W_{SS} \\ Y_D \end{bmatrix} \tag{A6}$$

is equal to the rank of the column-augmented matrix. Also, the number of outputs must be less than or equal to the number of controls. If the number of outputs equals the number of controls and M is nonsingular, the solution is unique and is given by the following:

$$\begin{bmatrix} M_{XW} & M_{XY} \\ M_{UW} & M_{UY} \end{bmatrix} = \begin{bmatrix} (A_D - I) & B_D \\ C & 0 \end{bmatrix}^{-1} \tag{A7}$$

The controller Eq. (27) can now be rewritten to account for the steady offsets:

$$U_K = U_{SS} - C_X[X_K - X_{SS}] \tag{A8}$$

which becomes the following:

$$U_K = -C_X X_K + C_Y Y_D - C_W W_{SS} \tag{A9}$$

where $C_Y = C_X M_{XY} + M_{UY}$; and $C_W = [C_X M_{XW} + M_{UW}]B_{WD}$. Note that so far we have assumed X_K and W_{SS} are known exactly. The method for actually estimating these quantities is discussed in Appendix B.

Linear Quadratic Gaussian Controller for Exponential Set Points

It is sometimes desirable to follow a reference input that is decaying exponentially in time. For example, during flare, the aircraft is asymptotically approaching an elevation below the runway. To create a controller for this situation, we simply modify the design procedure just described to follow a continuous desired output

$$Y_D(t) = Y_i e^{-at} \tag{A10}$$

which corresponds the following discrete reference command:

$$Y_{DK} = Y_i r^K \tag{A11}$$

where $r = e^{-aT}$, and T is the sample time.

Nominally, while tracking this reference command input,

$$X_K = X_i r^K \tag{A12}$$

$$U_K = U_i r^K \tag{A13}$$

Combining this with Eqs. (18) and (19), we have the following:

$$X_i \; r^{K+1} = A_D X_i r^K + B_D U_i r^K + B_{WD} W_K \tag{A14}$$

We simplify as before to find the following:

$$-B_{WD} W_K r^{-K} = [A_D - rI] X_i + [B_D] U_i \tag{A15}$$

$$Y_i r^K = [C] X_i + [0] U_i \tag{A16}$$

and

$$\begin{bmatrix} X_i \\ U_i \end{bmatrix} = \begin{bmatrix} N_{XW} & N_{XY} \\ N_{UW} & N_{UY} \end{bmatrix} \begin{bmatrix} -B_{WD} W_K r^{-K} \\ Y_D \end{bmatrix} \tag{A17}$$

where

$$\begin{bmatrix} N_{XW} & X_{XY} \\ N_{UW} & N_{UY} \end{bmatrix} = \begin{bmatrix} (A_D - rI) & B_D \\ C & 0 \end{bmatrix}^{-1} \tag{A18}$$

The controller equation can finally be written to account for exponential reference inputs:

$$U_K = U_{SS} - C_X [X_K - X_{SS}] \tag{A19}$$

which becomes

$$U_K = -C_X X_K + C_Y Y_{DK} - C_W W_K \tag{A20}$$

where

$$C_Y = C_X N_{XY} + N_{UY}$$

$$C_W = [C_X N_{XW} + N_{UW}] B_{WD}$$

We have again assumed X_K and W_K are known exactly. The method for estimating these quantities is discussed in Appendix B.

Integral Control Law Controller for Constant Set Points and Steady Wind Disturbances

The continuous form of the modified integral control was developed and presented by Holley and Bryson.[5] The same logic is followed to develop the discrete-time version.

One way to develop a controller that allows for nonzero set points and steady disturbance inputs is to define a general control law that will eventually lead to a form of integral control:

$$U_K = -C_X X_K + C_Y Y_0 \tag{A21}$$

Combining this with Eqs. (18) and (19), it turns out that if $[I - A_D + B_D C_X]$ is nonsingular, the steady-state solution is as follows:

$$X_{SS} = [I - A_D + B_D C_X]^{-1}[B_D C_Y Y_0 + B_{WD} W_{SS}] \tag{A22}$$

$$Y_D - TB_{WD} W_{SS} = TB_D C_Y Y_0 \tag{A23}$$

where $T \equiv C[I - A_D + B_D C_X]$.

Equation (A23) is satisfied for all Y_D and W_{SS} if

$$TB_D C_Y = I \tag{A24}$$

$$Y_0 = Y_D - TB_{WD} W_{SS} \tag{A25}$$

Solving for C_Y is straightforward when the number of outputs is equal to the number of controls and the square matrix (TB_D) is full rank. In order to compute Y_0, i.e., the implicit influence of the disturbances W_K in the control law, define a new state as follows:

$$Y_{WK} \equiv TB_{WD} W_K \tag{A26}$$

and define a new estimator for this state with gain L_2 (which is chosen empirically), so that

$$Ye_{W(K+1)} = Ye_{WK} - L_2[Ye_{WK} - Y_{WK}] \tag{A27}$$

Using Eq. (18)

$$Ye_{W(K+1)} = Ye_{WK} - L_2 TB_{WD} We_K + L_2 T[X_{K+1} - A_D X_K - B_D U_K] \tag{A28}$$

By applying Eqs. (A21), (A24), and (A25), Eq. (A28) can eventually be simplified to the following useful form:

$$Ye_{W(K+1)} - L_2 TX_{K+1} = Ye_{WK} - L_2 TX_K - L_2[Y_D - CX_K] \tag{A29}$$

This can be simplified even further by defining V_K so that

$$L_2 V_K = Ye_{WK} - L_2 TX_K \tag{A30}$$

Equations (A29) and (A30) become the following:

$$V_{K+1} = V_K - [Y_D - CX_K] \tag{A31}$$

We now have an expression that represents the integral of the error between desired output and actual (or measured) output. The control law can be determined by expanding Eq. (A21):

$$U_K = -C_X X_K + C_Y[Y_D - Ye_{WK}] \tag{A32}$$

$$U_K = -C_X X_K + C_Y[Y_D - L_2 V_K + L_2 TX_K] \tag{A33}$$

and finally

$$U_K = -[C_{XX}]X_K + [C_Y]Y_D - [C_V]V_K \tag{A34}$$

where

$$C_{XX} = C_X + C_Y L_2 T$$

$$C_Y = (TB_D)^{-1}$$

$$C_V = C_Y L_2$$

Note that the discrete form of the integral control law is similar to the continuous form described by Holley and Bryson.[5]

The size of L_2 is linearly related to the amount of control used to zero the integral error. The selection of L_2 for these simulations was done empirically using the following considerations: it must be large enough to produce a satisfactory response to disturbances and unmodeled errors, but cannot be so large that unacceptable control authority is required.

Integral Control Law Controller for Exponential Set Points and Wind Disturbances

The previous ICL controller design assumes reference inputs are constant and disturbances have a constant bias combined with zero-mean noise. To account for the exponential decaying input associated with the flare phase, the previous controller design is modified with the following assumptions:

$$Y_{DK} = Y_i r^K \tag{A35}$$

and nominally

$$X_K = X_i r^K \tag{A36}$$

With these assumptions made, the design for the new integral control law is almost the same as for the old one. We begin with the general control law:

$$U_K = -C_X X_K + C_Y Y_0 \tag{A37}$$

Combining this with Eqs. (18) and (19), it turns out that if $[rI - A_D + B_D C_X]$ is nonsingular, the flare solution is as follows:

$$X_i = [rI - A_D + B_D C_X]^{-1}[B_D C_Y Y_0 + B_{WD} W_K] \tag{A38}$$

$$Y_D - TB_{WD} W_K = TB_D C_Y Y_0 \tag{A39}$$

where

$$T \equiv C[rI - A_D + B_D C_X]^{-1}$$

Note that this T is slightly different from the T defined for the constant reference input case because of the r term.

Equation (A39) is satisfied for all Y_D and W_K if

$$TB_D C_Y = I \tag{A40}$$

$$Y_0 = Y_D - TB_{WD} W_K \tag{A41}$$

Solving for C_Y is straightforward when the number of outputs is equal to the number of controls and the square matrix (TB_D) is full rank. To compute Y_0, i.e., the implicit influence of the disturbances in the control law, define a new state as follows:

$$Y_{WK} \equiv TB_{WD}W_K \tag{A42}$$

and define a new estimator for this state with gain L_2 so that

$$Ye_{W(K+1)} = rYe_{WK} - L_2[Ye_{WK} - Y_{WK}] \tag{A43}$$

Using Eq. (18)

$$Ye_{W(K+1)} = rYe_{WK} - L_2TB_{WD}We_K + L_2T[X_{K+1} - A_DX_K - B_DU_K] \tag{A44}$$

By applying Eqs. (A37), (A40), and (A41), Eq. (A44) can eventually be simplified to the following useful form:

$$Ye_{W(K+1)} - L_2TX_{K+1} = r[Ye_{WK} - L_2TX_K] - L_2[Y_D - CX_K] \tag{A45}$$

This can be simplified even further by defining V_K so that

$$L_2V_K = Ye_{WK} - L_2TX_K \tag{A46}$$

Equations (A45) and (A46) become the following:

$$V_{K+1} = rV_K - [Y_D - CX_K] \tag{A47}$$

We now have an expression that represents the integral of the error between desired output and actual (or estimated) output. The control law can be determined by expanding Eq. (A37):

$$U_K = -C_XX_K + C_Y[Y_D - Ye_{WK}] \tag{A48}$$

$$U_K = -C_XX_K + C_Y[Y_D - L_2V_K + L_2TX_K] \tag{A49}$$

and finally

$$U_K = -[C_{XX}]X_K + [C_Y]Y_D - [C_V]V_K \tag{A50}$$

where $C_{XX} = C_X + C_YL_2T$; $C_Y = (TB_D)^{-1}$; and $C_V = C_YL_2$.

Transition from Glide-Slope Phase to Flare Phase

In the LQG controller the phase transition from glide-slope to flare is relatively straightforward. The winds are estimated along the glide slope, so these state variables are held and used during flare. The controller switches an internal state variable from *d* to *h*, updates the internal model to include this new state, and recomputes controller and estimator gains. The new controller gains account for flare error specifications (see Table 3).

The procedure is basically the same for the ICL controller, however, because the winds are not directly being estimated, care must be taken to update the integral states. The governing equations for the integral state for glide-slope phase and flare phase are, respectively:

$$V_{gsK} = L_{2gs}^{-1}T_{gs}B_{WD}We_K - T_{gs}Xe_{gsK} \tag{A51}$$

$$V_{flK} = L_{2fl}^{-1}T_{fl}B_{WD}We_ir^K - T_{fl}Xe_{flK} \tag{A52}$$

To match correctly at transition, V_{fl0} should be chosen so that the state and wind estimates match those reflected in V_{gsKf}. Equation (A51) can be expressed with the unknown quantity (We_K) as a function of the "known" quantities (V_{gsK} and Xe_{gsK}),

$$T_{gs} B_{WD} We_K = L_{2gs}[V_{gsK} + T_{gs} Xe_{gsK}] \tag{A53}$$

Because the solution for $B_{\mathrm{WD}} We_K$ is underdetermined, a least-norm solution can be found by using the pseudoinverse of T_{gs}:

$$B_{WD} We_K = T_{gs}^T (T_{gs} T_{gs}^T)^{-1} L_{2gs}[V_{gsK} + T_{gs} Xe_{gsK}] \tag{A54}$$

Substituting this into Eq. (A52), we get a solution for the integral error term at the beginning of flare in terms of the integral error term at the end of the glideslope phase and other known quantities:

$$V_{fl0} = L_{2fl}^{-1} T_{fl} T_{gs}^T (T_{gs} T_{gs}^T)^{-1} L_{2gs}[V_{gsKf} + T_{gs} Xe_{gsKf}] - T_{fl} Xe_{fl0} \tag{A55}$$

Appendix B: Discrete Time Optimal Estimator

Estimator Synthesis

In our design of the ICL and LQG controllers, we have assumed that the quantitites X_K and possibly W_K are available. In truth, these quantities must be estimated from the plant model, control history, and measurements. These estimates are the values actually used in the control law.

Integral Control Law Estimator

One way to implement the ICL is to assume the wind disturbances and sensor measurements are Gaussian, zero mean, and uncorrelated. We expect any biases caused by modeling errors and wind biases to be accounted for with the integral error term. Referring to Eqs. (18) and (20), we can write these assumptions in the following way:

$$E\{W_K\} = 0 \qquad \text{for all } K \tag{B1a}$$

$$E\{W_J W_K^T\} = W_D \qquad \text{for } J = K \tag{B1b}$$

$$= 0 \qquad \text{for } J \neq K \tag{B1c}$$

$$E\{\mu_K\} = 0 \qquad \text{for all } K \tag{B2a}$$

$$E\{\mu_J \mu_K^T\} = V_D \qquad \text{for } J = K \tag{B2b}$$

$$= 0 \qquad \text{for } J \neq K \tag{B2c}$$

$$E\{W_J \mu_K^T\} = 0 \qquad \text{for all } J, K \tag{B3}$$

The discrete covariance of the disturbance noise is W_D, and V_D is the discrete covariance of the measurement noise.

We can compute the prediction estimator gain L given W_D and V_D using standard linear quadratic estimation techniques.[1] We then generate state estimates using the following equations:

Measurement update:

$$X_K^* = Xe_K + L[Z_K - HXe_K] \tag{B4}$$

Time update:

$$Xe_{K+1} = A_D X_K^* + B_D U_K \tag{B5}$$

Control law:

$$U_K = -C_{XX} Xe_K + C_Y Y_{DK} - C_V V_K \tag{B6}$$

Linear Quadratic Gaussian Estimator

In the LQG case, we generate estimates of W_K in an attempt to directly compensate for wind disturbances. The straightforward approach of modeling these disturbances as constants is not helpful with a constant gain estimator. The optimal result is $We_{K+1} = We_K$, so the initial estimate is never changed as measurements are taken.

A more useful model is to assume the disturbance is exponentially correlated with time constant τ_w, which is long compared to the characteristic times of the aircraft. The continuous disturbance model is given by the following:

$$\mathrm{d}W(t)/\mathrm{d}t = -1/\tau_W W(t) + 1/\tau_W r(t) \tag{B7}$$

where

$$E\{r(t)\} = 0 \tag{B8}$$

$$E\{r(t) r^T(t)\} = W_C \delta(t) \tag{B9}$$

The size of W_C and τ_W determine how quickly the estimator responds to changes in wind. For a small W_C, the estimator reacts sluggishly in the presence of gusts; whereas, for a large W_C, the estimator reacts quickly, but performance is degraded because of measurement noise.

The disturbance vector is appended to the state vector before discretization. An estimator is designed using linear .quadratic techniques on the augmented state vectors and matrices to find an estimator for the plant states and disturbance states simultaneously.

Appendix C: Numerical Values for Continuous System

This appendix contains the continuous truth matrices used in the autopilot simulation. The numerical values were derived from Bryson[2], and are in units of feet, seconds, and centiradians.

Automatic Landing System—Longitudinal Mode

Matrix A:

$$\begin{matrix}
-0.0210 & 0.1220 & 0.0000 & -0.3220 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\
-0.2090 & -0.5300 & 2.2100 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0440 & 0.0000 \\
0.0170 & -0.1640 & -0.4120 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5440 & 0.0000 \\
0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0349 & -0.9994 & 0.0000 & 2.2087 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0349 \\
0.0000 & -1.0000 & 0.0000 & 2.2100 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.2500 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000
\end{matrix}$$

Matrix B:

$$\begin{matrix}
0.0100 & 0.0000 \\
-0.0640 & 0.0000 \\
-0.3780 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.2500 \\
0.0000 & 0.0000
\end{matrix}$$

Matrix B_W:

$$\begin{matrix}
0.0210 & -0.1220 \\
0.2090 & 0.5300 \\
-0.0170 & 0.1640 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000 \\
0.0000 & 0.0000
\end{matrix}$$

Glide-slope output distribution matrix C:

$$\begin{matrix}
0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000
\end{matrix}$$

Glide-slope measurement distribution matrix H:

$$\begin{matrix}
1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000
\end{matrix}$$

Wind disturbances:

Wind gradient:

0.1000 (ft/s)/ft

Steady wind:

25.000 ft/s

Automatic Landing System—Lateral Mode

Matrix A:

$$\begin{matrix} -0.0890 & -2.1900 & 0.0000 & 0.3190 & 0.0000 & 0.0000 \\ 0.0760 & -0.2170 & -0.1660 & 0.0000 & 0.0000 & 0.0000 \\ -0.6020 & 0.3270 & -0.9750 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.1375 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0094 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.0000 & 0.0000 & 0.0000 & 0.3010 & 2.1894 & 0.0000 \end{matrix}$$

Matrix B:

$$\begin{matrix} 0.0000 & 0.0327 \\ 0.0264 & -0.1510 \\ 0.2270 & 0.0636 \\ 0.0000 & 0.0000 \\ 0.0000 & 0.0000 \\ 0.0000 & 0.0000 \end{matrix}$$

Matrix B_W:

$$\begin{matrix} 0.0890 \\ -0.0760 \\ 0.6020 \\ 0.0000 \\ 0.0000 \\ 0.0000 \end{matrix}$$

Glide-slope output distribution matrix C:

$$\begin{matrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{matrix}$$

Glide-slope measurement distribution matrix H:

$$\begin{matrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{matrix}$$

Wind disturbances:

Wind gradient:

0.10000 (ft/s)/ft

Steady wind:

25.000 ft/s

Bibliography

Beser, J., and Parkinson, B. W., “The Application of NAVSTAR Differential GPS in the Civilian Community,” *Navigation*, Vol. 29, No. 2, 1982.

Etkin, B., “Dynamics of Atmospheric Flight,” Wiley, New York, 1972.

Parkinson, B. W., and Fitzgibbon, K. T., “Aircraft Automatic Landing System Using GPS,” *Navigation*, Vol. 42, No. 1, Jan. 1989.

Stengel, R. F., *Stochastic Optimal Control*, Wiley, New York, 1986.

References

[1] Bryson, A. E., and Ho, Y. C., *Applied Optimal Control,* Hemisphere, Bristol, PA, 1975.

[2] Bryson, A. E., *Control of Spacecraft and Aircraft,* Princeton University Press, Princeton, NJ, 1994.

[3] Bryson, A. E., and Henrikson, L. J., “Estimation using Sampled Data Containing Sequentially Correlated Noise,” *Journal of Spacecraft and Rockets,* Vol. 5, No. 6, 1982, pp. 662–665.

[4] Gelb, A., *Applied Optimal Estimation,* MIT Press, Cambridge, MA, 1974.

[5] Holley, W. E., and Bryson, A. E., “MIMO Regulator Design for Constant Disturbances and Non-zero Set Points with Applications to Automatic Landing in a Crosswind,” Stanford University, SUDAAR No. 465, Aug. 1973.

[6] Roskam, J., *Airplane Flight Dynamics and Automatic Flight Control,* The University of Kansas, Lawrence, KS, 1979.

Chapter 15

Precision Landing of Aircraft Using Integrity Beacons

Clark E. Cohen,* Boris S. Pervan,† H. Stewart Cobb,†
David G. Lawrence,† J. David Powell,‡ and Bradford W. Parkinson§
Stanford University, Stanford, California 94305

LANDING aircraft in poor visibility imposes the very highest standards of performance for a navigation system. Required to work under extreme weather conditions and at life-critical levels of performance, a Category III (lowest visibility) landing system must meet a vertical position accuracy requirement of 2 ft (95%) with extremely demanding integrity. For each approach, the probability of missed detection of failure cannot exceed 5×10^{-9}. This chapter explores the augmentation of GPS with *Integrity Beacons*—a special type of pseudolite—to achieve the required navigation performance (RNP) for precision landing of aircraft.

I. Overview of the Integrity Beacon Landing System

The Integrity Beacon Landing System (IBLS)[1–3] is illustrated in Fig. 1. It is founded on using GPS augmented with Integrity Beacons—compact, low-power, ground-based marker beacon "pseudolites" (transmitters used as pseudo-GPS satellites). Integrity Beacons are nominally situated in pairs on either side of the approach path to a runway. The power is set low so that the broadcast signal is measurable only inside of the "bubble" shown in Fig. 1. The bubble radius (determined by the broadcast signal power) is adequate when it is only a few times larger than the nominal altitude of approach. A conventional differential GPS (DGPS) reference station is located at the airport tower. This station broadcasts GPS reference information to all aircraft in the vicinity of the airport, both on the ground and in flight. Flying through the integrity bubbles, an aircraft is capable of tracking enough ranging sources to initialize DGPS to centimeter-level accuracy with a high degree of integrity. The aircraft can then maintain this initialization from bubble exit through touchdown and rollout.

[*] Research Associate, Department of Aeronautics and Astronautics; Manager, GPS Precision Landing.

[†] Ph.D. Candidate, Department of Aeronautics and Astronautics.

[‡] Professor, Department of Aeronautics and Astronautics.

[§] Professor, Department of Aeronautics and Astronautics; Director of GPS Program.

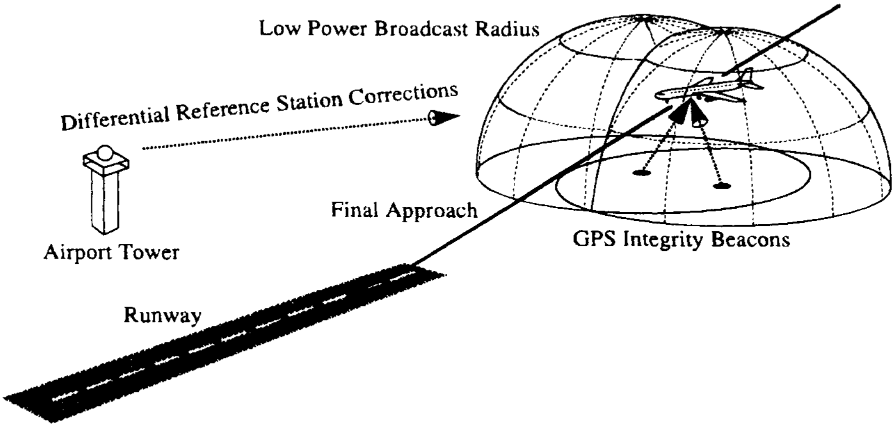

Low Power Broadcast Radius
Differential Reference Station Corrections
Airport Tower
Final Approach
GPS Integrity Beacons
Runway

Fig. 1 GPS Integrity Beacons for Category III precision landing.

A. Centimeter-Level Positioning

Kinematic aircraft positioning is based on precise measurements of the GPS carrier phase. The GPS L_1 carrier wavelength is 19 cm, and a state-of-the-art GPS receiver can measure the carrier phase in real-time to a small fraction of a wavelength (i.e., subcentimeter precision). Precision positioning is accomplished by measuring the carrier phase difference between an antenna at a surveyed location on the ground and the aircraft antenna. By resolving the carrier-phase cycle ambiguities (the number of integer wavelengths between each given pair of antennas in the direction of each given GPS satellite), a receiver can determine its position to centimeter-level accuracy.

Historically, cycle ambiguities have been resolved by using integer search techniques based on redundant ranging measurements with the optional use of the dual-frequency GPS signal.[4,5] In contrast, IBLS resolves integers by using the ranging information from the Integrity Beacons over a large change in geometry caused by aircraft motion. This allows the user to solve explicitly and analytically for the exact numerical values of the integers.

Initially flying in on traditional differential GPS, an IBLS-equipped aircraft flies over a pair of Integrity Beacons to resolve the cycle ambiguities. Thereafter, centimeter-level positioning accuracy is achieved, all the way through landing and taxi. The system utilizes the single-frequency C/A-code and carrier signal that has been explicitly provided for civilian use. Real-time operation is essentially independent of selective availability (SA).

By allowing the aircraft to use the precision of GPS carrier phase reliably Integrity Beacons yield centimeter-level sensor accuracy for the aircraft. The aircraft receiver can convert this accuracy into a high-level of onboard integrity. As it flies through the bubble, the aircraft obtains GPS ranging information from every direction—both from the GPS satellites in the sky above and from the Integrity Beacons below. If any element of the system is not performing to specification, the inconsistencies between measurements (precise to the centimeter level) make the problem clear, and the system issues an integrity alarm.

B. History of the Integrity Beacon Landing System

IBLS is a spin-off of NASA-sponsored research at Stanford University directed toward a satellite test of Einstein's General Theory Relativity. On this spacecraft, called Gravity Probe B, GPS will be used for both precise orbit determination and spacecraft attitude determination. A new high-performance attitude determination system based on GPS carrier phase was developed and flight tested on both spacecraft and aircraft.[6] Many of the kinematic positioning techniques pioneered in the attitude system laid the groundwork for the landing system. Under FAA sponsorship, the IBLS "Pathfinder" was developed as a feasibility test bed for Category III precision landing.

C. Doppler Shift and Geometry Change

The principle by which cycle ambiguities are resolved is similar to the familiar changing pitch of a passing locomotive whistle as heard by a stationary listener. As it flies overhead, the moving aircraft measures the carrier phase (the derivative of which is the Doppler shift) of the stationary ground-based transmitter signal. A large change in angular geometry occurs on a time scale of seconds. When referenced to the slowly changing satellite geometry, the Integrity Beacon carrier phase range measurements coupled with the large change in angular geometry quickly provide enough information to pinpoint the cycle ambiguities for each satellite.

A *single* Integrity Beacon below provides enough information to resolve altitude and along-track position (as described further in Sec. IV.A). Flying between a *pair* of Integrity Beacons placed on either side of the ground track (as in Fig. 1A) provides enough information to initialize all three components of aircraft position to high accuracy and also provide a crosscheck.

The code phase component of the GPS signal is not explicitly required for IBLS positioning. The only indirect requirements for code modulation are to enable the receiver to distinguish between different GPS satellite carrier signals and to provide a coarse position initialization for the algorithm. Therefore, because code-based ranging can be considered optional, it can provide an *additional, independent* layer of integrity checking. Of course the modulation can also include data communications in such a way that is similar to the satellite data messages.

II. Required Navigation Performance

Required navigation performance for precision landing is being quantified by the parameters *accuracy, integrity, availability,* and *continuity.* The IBLS performance in the context of these required performance parameters is discussed in the following subsections.

A. Accuracy

The consistent and dependable centimeter-level accuracy provided by the use of the GPS Integrity Beacons exceeds both the ICAO Annex 10[7] and Federal Radionavigation Plan (FRP)[8] system specification of 2-ft vertical (95%) navigation sensor error (NSE) for Category III landings with a substantial margin.

Another proposed specification—the RNP tunnel concept[9]—sets a 15-ft (95%) requirement at 100-ft altitude for total system error (TSE). At any given instant, TSE is the sum of NSE and the pilot or autopilot's flight technical error (FTE). In flight testing with different autopilots, IBLS has also met this specification.

The accuracy provided by the Integrity Beacon architecture has important implications in engineering the landing system to be resistant to adverse conditions. An important overall contributor to the utility of the Integrity Beacon may be its very low NSE. As discussed in Sec. V, Integrity Beacon positioning is largely insensitive to position dilution of precision (PDOP). The system can easily handle worst-case satellite failures and still maintain excellent NSE. Newer pilot displays may allow landing systems to meet the total accuracy specification at lower cost if larger pilot-in-the-loop FTE is tolerable. When NSE is small, the TSE becomes insensitive to NSE. On average, TSE is the root-sum-square of the two 95% components:

$$\mathrm{TSE} = \sqrt{\mathrm{NSE}^2 + \mathrm{FTE}^2} \cong \mathrm{FTE}$$

High sensor accuracy and the largest possible allowance for FTE also translates into more margin to safely reject unpredictable wind gust disturbances on final approach.

Very importantly, this ample margin for navigation sensor accuracy can be used to improve system integrity. As described next, the high accuracy means that extremely tight thresholds can be set for Receiver Autonomous Integrity Monitoring (RAIM).

B. Integrity

Integrity is the measure of trust that can be placed on the correctness of the navigation system output. The requirement for Category III integrity[7] is given as a probability of missed failure detection per approach of 5×10^{-9}.

Perhaps the most powerful benefit provided by Integrity Beacons is the capacity for RAIM during precision approach and landing. A precision approach position solution based on GPS Integrity Beacons is overdetermined. Because of the redundancy of information and the centimeter-level precision of the measurements, tight thresholds on the solution rms residual (typically on the order of tens of centimeters) can be set for the detection of anomalous conditions.

Receiver autonomous integrity monitoring provides an important improvement over the traditional Ground Monitor Station used for integrity. Perhaps the most direct benefit of RAIM is that it covers failure modes in *all* segments of the system, including the aircraft segment. Ground monitors have no way of resolving these types of failures. Another problem with traditional ground integrity monitoring schemes alone [even with a landing system as mature as instrument landing systems (ILS)] is that there are still some nonaircraft error modes that can still slip through undetected. In the case of the GPS, one such potentially dangerous error mode is that recently associated with space vehicle pseudo random noise code (PRN) 19.[10] In the case of PRN 19, a GPS signal that was somewhat abnormal affected receivers from different manufacturers in different ways over a period of a few months. Some receivers experienced range biases that differed by several meters. Suppose that both the ground reference receiver and the ground integrity monitor were affected identically, while the aircraft receiver was affected differently. No integrity warning would be issued. However, a potentially dangerous situation for the aircraft could exist.

The IBLS (which provides both ground monitoring *and* RAIM) is immune to the class of PRN 19-type anomalies and other failure modes, because it employs the GPS carrier to solve explicitly for all range biases (see Sec. IV). An integrity detection scheme that emphasizes RAIM enables the ultimate integrity decision to be made by the aircraft, not the ground. This autonomous decision capability ensures that, regardless of the state of the system ground components (including any monitoring equipment), there is always more than enough information for the aircraft to make an independent assessment of integrity.

Overall system integrity is analyzed quantitatively in Sec. VII. Analysis indicates that a probability of missed detection of 10^{-9} in actual flight operations will be achievable through the use of Integrity Beacons.

C. Availability

Availability is the fraction of time that the complete landing system will be able to carry out its function at the initiation of the intended approach. Note that in this case, *availability* also implies the satellite constellation geometry for the availability of *integrity.*

Use of the GPS Integrity Beacons significantly augments the availability of GPS landing capability by providing additional ranging measurements. IBLS requires only four GPS satellites for full performance when an inertial reference unit (IRU) is employed (see Sec. VI). Therefore, the availability of IBLS is projected to be *better than that of enroute GPS navigation,* which requires more than four satellites for the redundancy needed to carry out RAIM. IBLS also provides significant margin against high PDOP (see Sec. V.C). These advantages provide important protection in scenarios where satellite failures reduce coverage over a large geographical area. In such scenarios, flying to an adjacent airport is not a practical alternative.

D. Continuity

Continuity is a measure of interruptions in the system operation once an approach has been initiated. Because of the high accuracy afforded by the GPS Integrity Beacon concept, the probability of RAIM integrity false alarms can be brought to 10^{-7} (the continuity of service requirement) *or lower* if desired. Section VII provides a further discussion on the issues involved.

It may be possible to rely simply on redundant GPS satellites to satisfy continuity of integrity, especially if direct ILS receiver replacement is adopted for retrofitting existing aircraft. In the long run, however, it may be more beneficial to optimize the system to take full advantage of the IRU that is typically employed to drive the control surfaces of the aircraft. In this case, the GPS Integrity Beacon can be used to initialize accurately the three-axis position biases of the inertial unit at a safe altitude, allowing the aircraft to continue on even in the event of jamming or a complete failure of GPS.

III. Integrity Beacon Architecture

The centerpiece of IBLS is the Integrity Beacon itself. Working in conjunction with a traditional local area differential GPS ground station, this low-power transmitter provides the required ground augmentation of GPS for an aircraft to carry out a high-integrity landing. In this section, two types of Integrity Beacons are described: the *Doppler Marker* and the *Omni Marker.*

A. Doppler Marker

The *Doppler Marker* Integrity Beacon is an independent, low-power GPS signal transmitter that interfaces directly to the differential reference station. The Doppler marker circuit board is shown in Fig. 2. Designed to be the size of a business card, this transmitter is capable of running for more than half a day on an ordinary 9-volt battery. For a 300-m radius bubble, the transmitted power is on the order of a microwatt (−30 dBm). This type of GPS Integrity Beacon was used for the flight test results presented in Sec. V.

Figure 3 shows a block diagram of the complete system using the Doppler Marker Integrity Beacon. The signals from the beacons are fed directly into the differential reference station. The reference station measures the carrier phase of both the Integrity Beacon signals and the GPS satellite signals. Both sets of measurements are transmitted as a group up to the aircraft via the traditional differential data link.

B. Omni Marker

The Omni Marker is a more advanced version of the Integrity Beacon that offers several improvements over the Doppler Marker. When applied to IBLS, the Omni Marker eliminates most of the need for cabling to connect the various components of the ground system. The concept is illustrated in Fig. 4. In one continuous signal-processing chain, the Omni Marker locks onto the GPS signal from the receive antenna, strips off the satellite PRN code, reapplies a new pseudolite PRN code to the carrier, and rebroadcasts the signal. The outgoing code and carrier are kept phase coherent with respect to their incoming counter-

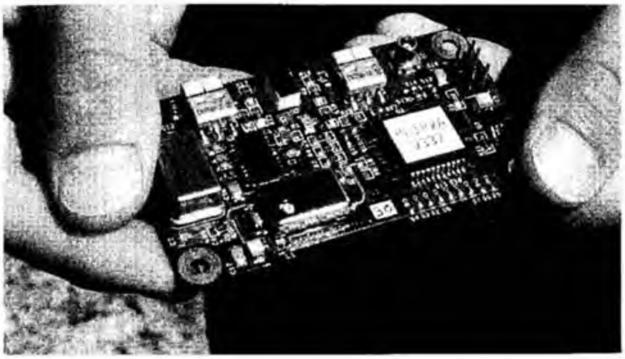


Fig. 2 GPS Integrity Beacon hardware.

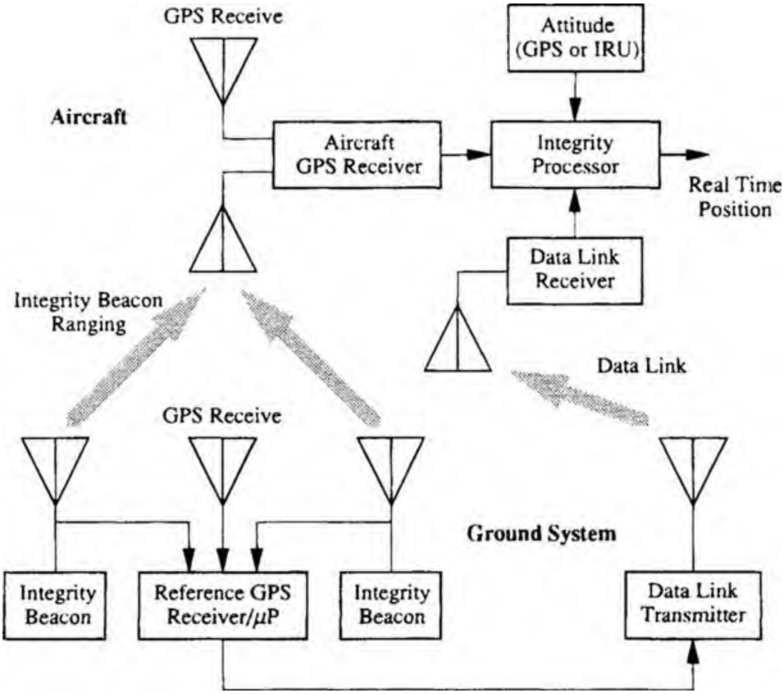


Fig. 3 Integrity Beacon Landing System diagram.

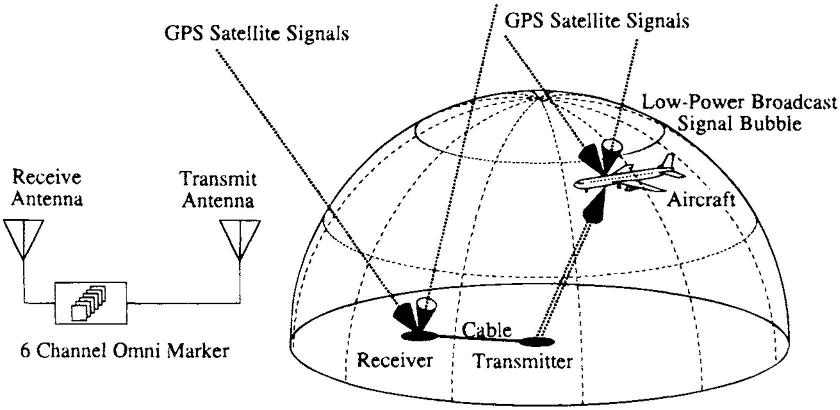


Fig. 4 Omni Marker Integrity Beacon concept.

parts. Ideally, the transmit and receive antennas would be collocated, but in practice some separation may be required for radio frequency isolation. Each Omni Marker Integrity Beacon is autonomous and independent. The result is a reliable, federated architecture that is much less vulnerable to individual component failures.

In addition to being the key to dependably resolving cycle ambiguities, the omni marker can also serve to relay the ground reference measurements to the aircraft. The digital data link traditionally used in DGPS (usually situated centrally at the airport) can be replaced by a coherent rebroadcast of all the received GPS signals. This application is further described in Ref. 71.

IV. Mathematics of Cycle Resolution

This section presents the mathematics of cycle ambiguity resolution using Integrity Beacons. Section IV.A on observability analysis offers a qualitative description of how Integrity Beacons are used and a look at what components of position can be resolved. Section IV.B on matrix formulation offers a quantitative description of how IBLS is able to carry out precision positioning and achieve high integrity.

A. Observability Analysis

A simplified analysis illustrates how a single Integrity Beacon provides both altitude (radial) and along-track position. Figure 5 shows a simple linear trajectory directly over the Integrity Beacon, located at the origin. The aircraft coordinates are along-track position *x*, cross-track (lateral, into the page) position *y*, and constant altitude (radial) *z*. The magnitude of the aircraft position vector (*x*,*y*,*z*) is the range to the pseudolite *r*. The measured range φ is the single-difference carrier phase measured between the ground reference receiver and the aircraft receiver via its belly-mounted antenna:

$$\varphi \equiv \varphi_{\text{aircraft}} - \varphi_{\text{reference}} = r + b - \Delta t_{\text{aircraft}} + \Delta t_{\text{reference}}$$

where *b* represents the sum of the cycle ambiguity and all system and cable biases and delays for the Integrity Beacon ranging link, and each Δt represents each receiver clock bias. Initialized with a trial trajectory from differential GPS, conventional kinematic positioning is used to eliminate the relative clock bias ($\Delta t_{aircraft} - \Delta t_{reference}$) between the two receivers. The Integrity Beacon range can then be measured directly, subject to the additive bias *b*.

$$\varphi' \equiv \varphi + (\Delta t_{\text{aircraft}} - \Delta t_{\text{reference}}) = r + b = \sqrt{x^2 + y^2 + z^2} + b$$

Linearizing the measured phase about the nominal trajectory, which (nominally)

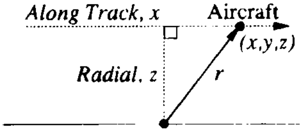


Fig. 5 Overflight geometry.

runs directly over the pseudolite ($y = 0$), it can be shown that the nominal observation matrix is given by the following:

$$d\varphi'|_{y=0} = \begin{bmatrix} \frac{x}{r} & 0 & \frac{z}{r} & 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \\ dz \\ db \end{bmatrix}$$

For reference, these observation functions (i.e., the row vector in the above equation) are plotted in Fig. 6 (normalized by the radius of closest approach) as a function of along track position. As long as the ranging signal is observed over a large enough arc, each observable component of the Integrity Beacon ranging signal is clearly distinguishable, including the most important parameter, altitude.

Note that the cross-track (lateral) component of position y is unobservable with a simple, linear trajectory over the Integrity Beacon. *For this reason, the dual Integrity Beacon configuration of Fig. 1 is used.* With two transmitters on either side of the glide slope, all three components of position are directly observable. In many cases, it may also be possible to use the same information from another Integrity Beacon placed under the glide slope of a parallel runway.

B. Matrix Formulation

To provide further insight as to how IBLS is able to provide such high performance accuracy and integrity using the GPS, the matrix formulation of cycle ambiguity resolution is presented here. The mathematical development of the system is defined with respect to Fig. 2, which shows a block diagram of the flight test system. The development of cycle ambiguity resolution algorithms is most easily done within the context of conventional carrier-based differential ranging with the Doppler marker, although the development is readily adaptable to the omni marker Integrity Beacon. Figure 7 serves as a guide for the vector definitions employed herein.

Single differencing of raw carrier phase measurements obtained at airborne and reference station receivers yields for the space vehicle (SV) i at epoch k the following:

$$\varphi_{ik} = -\hat{s}_{ik}^T \boldsymbol{x}_k + \tau_k + N_i^s + \epsilon_{ik}^s$$

where φ_{ik} is the single-differenced (aircraft minus reference) SV phase; $\hat{s}_{ik}$ is the

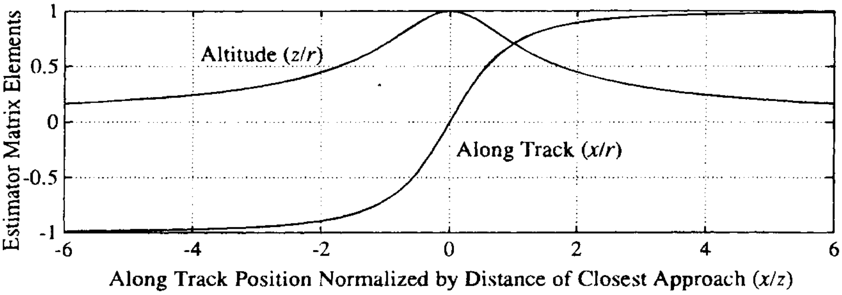


Fig. 6 Error profile for a pseudolite pass.

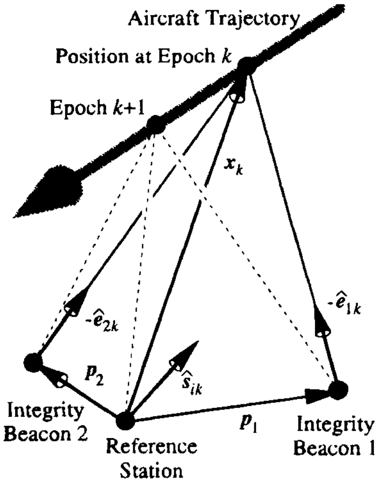


Fig. 7 Vector geometry.

line-of-sight unit vector to the SV; x_k is the displacement vector from the differential station GPS receive antenna to the top-mounted aircraft GPS antenna; τ_k is the difference in the aircraft and reference receiver clock biases; N_i^s is the satellite integer cycle ambiguity; and ϵ_{jk}^s is the satellite range measurement error caused by multipath and receiver noise. Similarly for Integrity Beacon j at epoch k, we have the following:

$$\phi_{jk} = |p_j - x_k| - |p_j| + \tau_k + N_j^p + \epsilon_{jk}^p$$

where ϕ_{jk} is the single-differenced Integrity Beacon phase, and p_j is the vector from the differential station to Integrity Beacon j. Because the transmitter is quite close, the formulation for satellites (whose wave fronts are essentially planar) is not appropriate. Instead, the use of range magnitude is necessary.

Given an approximate trajectory $\bar{x}_k$ obtained from code-based DGPS, the equations above can be expressed in terms of the deviation from the approximate trajectory: $\delta x_k \equiv x_k - \bar{x}_k$. Keeping first-order terms only, the result is as follows:

$$\delta\varphi_{ik} \equiv \varphi_{ik} + \hat{s}_{ik}^T\bar{x}_k = -\hat{s}_{ik}^T\delta x_k + \tau_k + N_i^s + \epsilon_{ik}^s$$

and

$$\delta\phi_{jk} \equiv \phi_{jk} - |p_j - \bar{x}_k| + |p_j| = -\hat{e}_{jk}^T\delta x_k + \tau_k + N_j^p + \epsilon_{jk}^p$$

where $\hat{e}_{jk}^T \equiv (p_j - \bar{x}_k)/|p_j - \bar{x}_k|$. To resolve cycle ambiguities, the value of one integer must be specified because of the existence of the clock bias τ_k, which is common to all measurements at epoch k. For simplicity, we choose $N_1^s = 0$. Defining $\delta\Phi_k$ to be the vector of m SV and two Integrity Beacon measurements at epoch k

$$\delta\Phi_k \equiv \begin{bmatrix} \delta\varphi_{1k} \\ \vdots \\ \delta\varphi_{mk} \\ \delta\phi_{1k} \\ \delta\phi_{2k} \end{bmatrix}, \qquad \text{and} \quad \hat{S}_k \qquad \text{as} \quad \hat{S}_k \equiv \begin{bmatrix} -\hat{\boldsymbol{s}}_{1k}^T & 1 \\ \vdots & \vdots \\ -\hat{\boldsymbol{s}}_{mk}^T & 1 \\ -\hat{\boldsymbol{e}}_{1k}^T & 1 \\ -\hat{\boldsymbol{e}}_{2k}^T & 1 \end{bmatrix}$$

we stack all n measurements collected during Integrity Beacon overpass to obtain the following:

$$\begin{bmatrix} \delta\Phi_1 \\ \vdots \\ \delta\Phi_k \\ \vdots \\ \delta\Phi_n \end{bmatrix} = \begin{bmatrix} \hat{S}_1 & 0 & \cdots & 0 & 0 & \bar{I} \\ 0 & \ddots & 0 & \ddots & 0 & \vdots \\ \vdots & \ddots & \hat{S}_k & \ddots & \vdots & \bar{I} \\ 0 & \ddots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & 0 & \hat{S}_n & \bar{I} \end{bmatrix} \begin{bmatrix} \delta\boldsymbol{x}_1^* \\ \vdots \\ \delta\boldsymbol{x}_k^* \\ \vdots \\ \delta\boldsymbol{x}_n^* \\ N \end{bmatrix} + \epsilon$$

where

$$\bar{I} = \begin{bmatrix} 0 & \cdots & 0 \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

$$\delta\boldsymbol{x}_k^* = \begin{bmatrix} \delta\boldsymbol{x}_k \\ \tau_k \end{bmatrix}$$

and

$$N = [N_2^s \quad \cdots \quad N_m^s \quad N_1^p \quad N_2^p]^T$$

The least-squares solution to the above can be obtained efficiently by sparse matrix batch algorithms or equivalently by sequential forward–backward smoothing. Because of the nonlinear nature of the problem, the "solution" $\delta\boldsymbol{x}_k$ is not the final answer. Instead, the approximate trajectory and observation matrix must be improved by the computed estimate of $\delta\boldsymbol{x}_k$, and the process above repeated through convergence (i.e., until the update $\delta\boldsymbol{x}_k$ becomes negligible). Computation time for convergence takes considerably less than a second on a 25-MHz 486 PC. Experience has shown that the solution converges in 3–10 iterations. The current algorithm has been tested in simulation and always converged when presented with an initial condition within 300m of the correct value for a 100m altitude bubble pass. In repetitive flight trials, presented in the following section, the algorithm has converged on every approach. In the unlikely event of convergence failure, the signal to the pilot would be a continuity alarm at 200 ft, *not* an integrity problem.

V. Experimental Flight Testing

This section quantifies the centimeter-level accuracy of the landing system and describes a sampling of the real-time flight testing that has occurred. Most of this section covers flight tests that use a laser tracker as the means for establishing navigation sensor accuracy. For the purposes of Category III precision landing, the approximate 1-ft accuracy of the laser tracker is satisfactory. However, it is claimed throughout this chapter that IBLS is capable of providing centimeter-level accuracy—an order of magnitude better than the laser tracker. The following introduction is provided to quantify these claims.

A. Quantification of Centimeter-Level Accuracy

Quantification of the centimeter-level accuracy of the system is indirect, because there is no other independent positioning sensor (including a laser tracker) known that is practical enough to facilitate comparison to the required level of performance. Therefore, a partial list of indirect means for establishing the centimeter accuracy of IBLS are given as follows:

1) Position checks against independent GPS static surveying.
2) Dynamic comparison of attitude determination using GPS against an IRU.

1. Position Checks Against Independent GPS Static Surveying

This method of comparison checks kinematic GPS initialized with Integrity Beacons against static survey with the GPS after the airplane lands. Each of these techniques measures position with respect to the defined reference runway coordinate frame. As presented in Sec. V.A, the quantitative agreement between these two independent means of positioning support absolute positioning accuracy to the centimeter level. In fact, centimeter-level static results are routinely obtained in surveying with satellite range rates in excess of 1 km/s.

2. Dynamic Comparison of Attitude Determination Using GPS Against an Inertial Reference Unit

This method of comparison checks relative positioning of antennas on an aircraft using GPS (employed primarily as a means of attitude determination) and the same quantity determined from an independent IRU. Chapter 19, Fig. 11 (this volume) shows an example of such relative positioning. Translating the 0.05 deg angular error of the IRU through a baseline of 16 m, the resulting dynamic position error is 1.4 cm rms. Translational experiments used to check GPS against an IRU position provide similar results over the short term, such as in Ref. 12.

These two comparison techniques—one static and absolute; the other dynamic and relative—combine to support the centimeter-level positioning accuracy using IBLS. Static survey results indicate the absolute accuracy of IBLS. Then, inertial comparison is one way to validate that kinematic survey techniques are just as accurate as static survey.

B. Piper Dakota Experimental Flight Trials

For the flight trials, the landing system hardware shown in Fig. 2 was set up in a single-engine Piper Dakota. Onboard the aircraft, signals from a Trimble TANS Quadrex receiver (specially modified for precision landing at Stanford University) were fed into the flight computer. A second GPS receiver, a TANS Vector, was used for attitude determination to supply the lever-arm correction for the positions of the belly-mounted Integrity Beacon receive antenna, landing gear, laser altimeter, and laser tracker retro reflector. Figure 8 is a photograph of the aircraft in flight. The fig. shows the four GPS antennas used for attitude determination mounted on the fuselage, tail, and each wingtip. Figure 9 shows a close-up of the integrity antenna mounted on the underside.

The relative positions of the ground station reference and Integrity Beacon transmit antennas are known to the subcentimeter level using standard GPS static surveying techniques. This procedure emphasizes one of the operational advantages of IBLS. The ground antennas may be placed wherever convenient. Then, they are self-surveyed with GPS. The system has proved it is ready to support the first flight inspection landing within an hour of the initial antenna placement.

By employing the simple hemispherical, upward-looking antenna pattern, site-specific multipath is of negligible consquence. Carrier multipath (which is approximately 1000 times smaller than code phase multipath) typically accounts for less than 0.5 cm of ranging error. Therefore, the flight inspection process is actually required only as a simple check of the installation database parameters and for obstacle clearance. Because Integrity Beacon antennas are upward-looking rather than side-looking, as with ILS, site-specific multipath is not an operational issue with GPS augmented with Integrity Beacons. Flight inspection should be simplified.

1. *Independent GPS Survey Results*

To demonstrate the centimeter-level accuracy of positioning using Integrity Beacons, position fixes from IBLS were compared with those from an independent GPS static survey. The cycle ambiguities were resolved in flight using Integrity Beacons broadcasting from the approach path at Palo Alto (CA) airport 1-km out. The aircraft flew an approach and landed without losing lock on the integers.

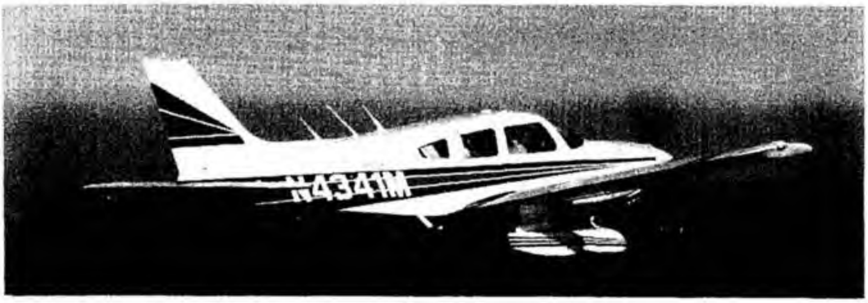


Fig. 8 Piper Dakota flight test aircraft.

Fig. 9 Belly-mounted integrity antenna.

After the aircraft was secured at the tie-downs, the final real-time position fix output by IBLS was recorded and compared to an independent estimate of this final position obtained using standard commercial static GPS survey receivers and software. These results are shown in Table 1. The quantity marked Δ is the difference between IBLS and the static survey. The quantity σ is the estimated standard deviation of Δ based on the covariance derived from the measurement geometry of the bubble pass. Repeated comparison experiments of this sort have shown that the static agreement is consistently on the centimeter level. Based on the discussion at the start of this section, it is believed that this same level of accuracy is available on a point-by-point basis throughout the entire portion of the flight following the bubble pass.

Table 1 Comparison of the Integrity Beacon Landing System and static survey

cm	Altitude	In-track	Cross-track
Δ	−0.1 cm	1.2 cm	1.4 cm
σ	1.7 cm	0.6 cm	1.1 cm

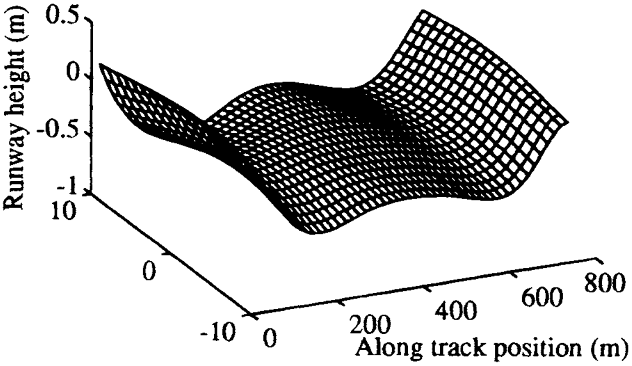


Fig. 10 Palo Alto Airport runway.

2. *Laser Altimeter Comparison*

For the laser altimeter tests, a laser rangefinder was installed in the aircraft.[2] To compare the laser altimeter data with IBLS position fixes, an accurate model of the Palo Alto Airport runway height was needed. Over a half-hour period, a comprehensive kinematic GPS survey was performed by driving a golf cart up and down the runway. A three-dimensional surface was fitted to these data to give a precise computerized model of runway height as a function of horizontal position.

Using 15 coefficients, the surface shown in Fig. 10 was generated. Very little memory storage is required to represent a runway in this manner range of the coefficients of 6 m, the total required data storage is only 15 bytes using fixed-point storage and a resolution of 5 cm. With this simple model, IBLS is able to output accurate height above the runway without the need for an extensive runway data base. The total measured discrepancy between the GPS and the laser altimeter for a series of seven touch-and-gos is summarized in Table 2.

3. *Laser Trackers*

To obtain real-time confirmation of IBLS accuracy in all three axes, a number of flight tests using laser trackers have been carried out. These include approaches with the Piper Dakota at the NASA Ames Crows Landing facility in August, 1993,[2] autocoupled approaches with an FAA Beech King Air at the FAA Technical Center in July and August, 1994,[11] and automatic landings of a United Airlines Boeing 737 at NASA Crows Landing in October, 1994.[13] The laser tracker

Table 2 Integrity Beacon Landing System and laser altimeter differences on Piper Dakota (7 approaches)

Mean	Standard deviation	Total estimated measures Error
3 cm	11 cm	< 30 cm

Table 3 Navigation sensor error at 50ft on King Air (49 approaches)

NSE, m	Vertical	Cross-track	Along-track
Sigma (σ)	0.2 m	0.1 m	0.2 m
Mean (μ)	−0.1 m	0.1 m	−0.2 m
$\vert\mu\vert + 2\sigma$ (95% error)	0.5 m	0.3 m	0.6 m

measures the azimuth, elevation, and range to the retroreflector mounted on the aircraft. The range accuracy is specified at ±0.3 m. Azimuth and elevation accuracy at the two facilities are specified at better than ±0.2 mrad 1σ. The laser tracker results have proved to be nearly identical for each flight trial, essentially bounding the accuracy of IBLS by that of the laser tracker. Furthermore, the accuracy is effectively independent of the type of aircraft employed. Representative data from the flight trials detail the laser-tracking results in the following subsections.

C. Federal Aviation Administration Beech King Air Autocoupled Approaches

In July and August, 1994, a series of 49 autocoupled approaches were carried out in an FAA Beech King Air at the FAA Technical Center. Initially the aircraft was guided using conventional code-based DGPS to bring it down the approach path over the Integrity Beacons at 600 ft. Upon bubble exit at roughly 500 ft, the system performed its cycle ambiguity calculation (in roughly 0.2 s using a Pentium processor) and assumed its precise-positioning mode. With the safety pilot monitoring the ILS, the pilot had the option to disengage the autopilot at 100 ft. However, because of the steady guidance being displayed and the smooth descent of the aircraft, the pilots typically left the autopilot engaged down to 50 ft or lower. On all 49 approaches, IBLS successfully resolved the cycle ambiguities, performed its internal onboard integrity checks, provided navigation output to within the accuracy of the laser tracker (or better), and autocoupled into the flight controls to guide the aircraft through the approach.

1. Navigation Sensor Error

Ensemble statistics for the NSE at 50-ft altitude are assembled in Table 3. Based on previous calibration experiments (see Independent GPS Survey Results, above), it is believed that most of the error is attributable to the laser tracker. The vertical NSE of 0.5 m meets the 0.6 m, 95% error requirement found in both ICAO Annex 10[7] and the Federal Radionavigation Plan (FRP). The ensemble statistics for cross-track NSE of 0.3 m exceeds both the ICAO 95% requirement of 4.4 m and the FRP 95% requirement of 4.1 m by a wide margin.

Table 4 shows statistics that summarize the accuracy achieved with respect to requirements.

2. Total System Error

Figure 11 shows the vertical TSE for the approaches. For comparison, the 95% inner tunnel boundaries[9] are superimposed on the plot. The TSE is plotted

Table 4 Integrity Beacon navigation sensor error performance on King Air (49 approaches)

95% Error, m	Vertical	Cross-track
Integrity beacon	0.5 m	0.3 m
ICAO Annex 10[7]	0.6 m	4.4 m
Federal Radionavigation Plan	0.6 m	4.1 m
Meets requirement	YES	YES

as a function of altitude for a nominal 3-deg glide slope. In other words, the along-track component of position is scaled to the nominal 3-deg glide slope altitude where the aircraft should be for each particular along-track position.

The plots begin at the bubble exit point, so the first part of the plot shows the small transient that occurs as the sensor accuracy increases. At the extreme right, the pilot breaks off the pass near the aim point and resumes manual control. Using a Category I autopilot, at no point does the TSE approach the 95% Category III inner-tunnel boundary. In other words, with the near-perfect sensor accuracy of IBLS, a Category I autopilot was capable of achieving Category III vertical TSE specifications. In the long run, the nearly perfect NSE of the Integrity Beacon may enable a future generation of safe, low-cost landing systems based on less expensive autopilots or pilot-in-the-loop graphical displays. These systems may be able to meet the TSE requirements by allowing for larger FTE.

The TSE results are summarized in Table 5 at 50-ft altitude. In spite of a significant cross-track bias in the autopilot, this Category I autopilot meets Category III specifications.

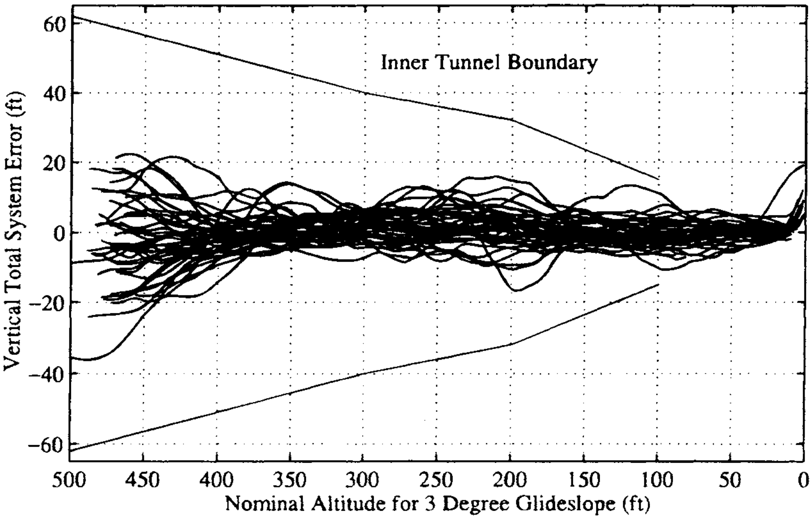


Fig. 11 Vertical total system error for Federal Aviation Administration King Air approaches.

Table 5 Integrity Beacon total system error at 50-ft altitude on King Air

Total system error, m	Vertical	Cross-track
Sigma (σ)	1.0 m	2.7 m
Mean (μ)	−0.5 m	5.3 m
$\lvert\mu\rvert + 2\sigma$ (95% error)	2.5 m	10.7 m
Required navigation performance tunnel	4.5 m (TBD)	15.5 m
Meets requirement	YES	YES

D. Automatic Landings of a United Boeing 737

In October, 1994, a United Airlines Boeing 737-300 was modified to accept IBLS guidance and was used to carry out 110 successful automatic landings using the GPS. Figure 12 shows the aircraft just following touchdown during one of its 110 autolands at NASA's Crows Landing research facility in California's Central Valley. These flight trials were sponsored by the FAA to help establish the feasibility of Category III precision landing using GPS.

The landing system configuration is shown in Fig. 13. A pair of standard ARINC 743 GPS antennas were mounted on fuselage—one on the top and and one on the underside of the aircraft. The Trimble TANS GPS receiver sent raw carrier-phase measurements to the navigation processor. These phase measurements—combined with the data link messages received through a VHF blade antenna on the top of the aircraft and attitude measurements from the inertial unit—provided the raw information for the IBLS flight computer to calculate precise aircraft position and glide path deviation. From the single-channel navigation processor, a dual-channel analog interface provided ILS localizer and glide slope signals to the autopilot. The 737-300 is equipped with a dual-channel flight control system designed for Category IIIA landings. The autolands were performed through touchdown without roll-out guidance.

A total of 111 approaches were attempted with 110 resulting in successful autolands. When the aircraft was at about 300 ft of altitude on the 37th approach (following the bubble pass), a U.S. Air Force upload transmission temporarily brought down the signal of one of the GPS satellites. As intended in the design, the landing system responded by raising a flag and calling off the approach

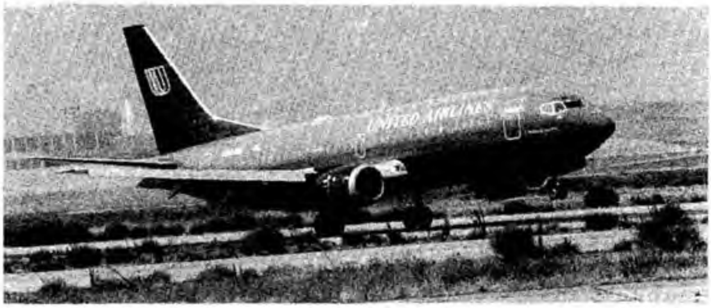


Fig. 12 One of 110 United Airlines Boeing 737 autolands.

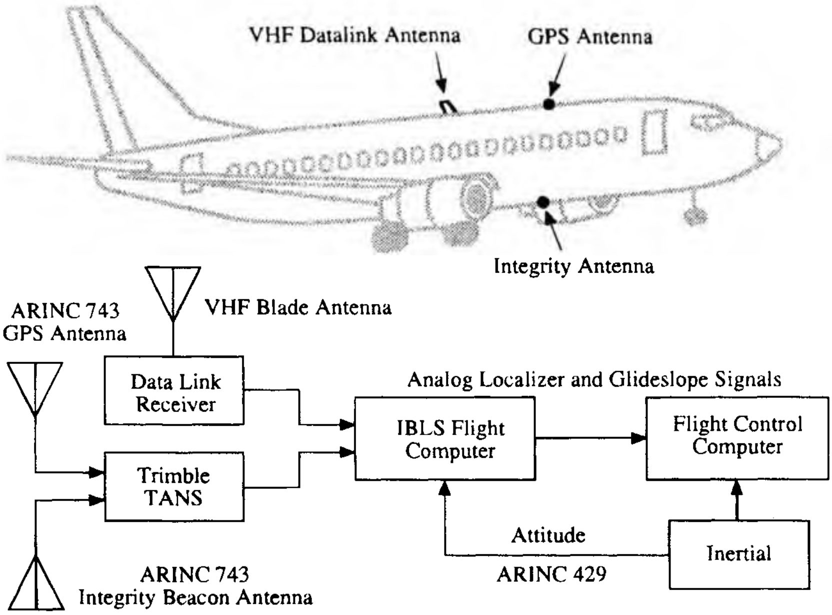


Fig. 13 Airborne configuration of United Boeing 737–300 Autoland Tests.

within 1/4 s of the event. The Category III specification for time-to-alarm caused by a system fault is 2 s. A second-generation system now allows landing to continue past such rare events using redundant GPS satellites or an inertial unit. Out of the total of 111 approaches flown, there were no false alarms and no missed detections.

1. Navigation Sensor Error

Figure 14 shows the vertical (most challenging) NSE for 100 of the autolands using the laser tracker as a reference. The plot is given as a function of distance to the aim point, converted into units of altitude assuming the standard 3-degree glide slope. To ensure that the approaches represented a true basis for operational evaluation, the plot shows only those 100 autolands for which the cycle ambiguities were intentionally reset (cleared) as a matter of procedure upon rollout onto final approach. (For experimentation purposes during some of the other autolands, it was demonstrated that the integers from a previous touch-and-go could be successfully carried around the pattern through to the next bubble pass.)

Again, it is believed that the error shown in Fig. 14 is dominated by the laser tracker. The standard error signature of the angular-based laser tracker is readily apparent in the plot as the spread on the vertical error increases with range. Prior to the advent of GPS, laser trackers have been traditionally considered the most accurate and convenient means of independently establishing position. It is interesting that GPS can be credited with finding new sources of error in laser trackers not previously considered nor encountered in this application. In the flight trials

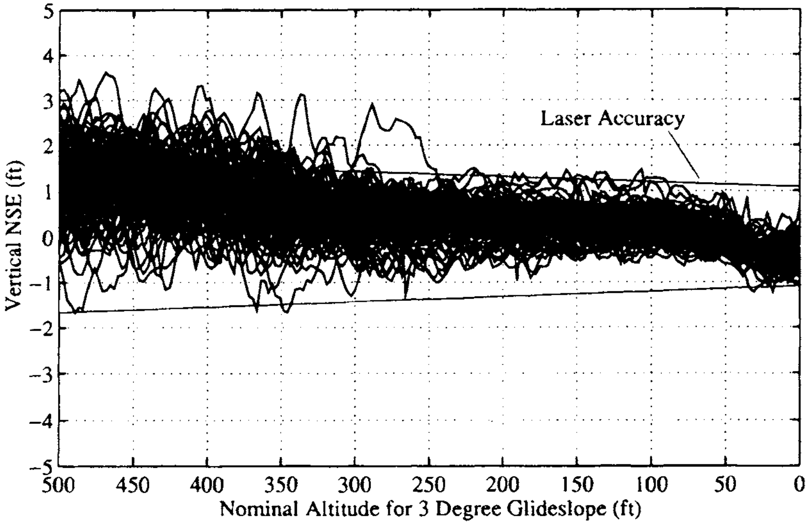


Fig. 14 Vertical navigation sensor error for 100 autolands.

at different test ranges, systematic errors found during data analysis on the order of tenths of a milliradian were traced to the omission of a correction for tropospheric refraction of the laser beam attributable to the gradient of atmospheric density with altitude. It also seems that other small systematic errors may still remain, such as harmonic noise in the elevation resolvers.

2. *Safety Margin Attributable to Enhanced Availability*

As a demonstration of the capacity of IBLS to provide enhanced availability, many of the approaches shown used a satellite selection algorithm picking the four highest elevation satellites in the sky. Occasionally this algorithm yields a PDOP greater than 10 for the four satellites. Interestingly, the positioning error in these high PDOP approaches is unnoticeable, because it is still outweighed by the laser error. In Fig. 14, even three autolands with satellite PDOP in the range of 17–18 do not stand out from the rest. In aircraft navigation applications with less-challenging performance requirements than Category III precision landing, a PDOP this large would be considered unusable. However, with IBLS, satellite PDOP of 18 resulted in no more than 1 ft of vertical error.

The statistical results confirm the high accuracy of the system. As expected, the statistics for NSE are comparable to those in Table 3. Table 6 summarizes

Table 6 Integrity Beacon flight technical error on the United Boeing 737-300

Flight technical error, m	Vertical, 100 ft	Cross-track, 100 ft	Vertical, 50 ft	Cross-track, 50 ft
Sigma (σ)	1.1 m	2.2 m	1.0 m	2.1 m
Mean (μ)	0.1 m	0.2 m	0.1 m	0.1 m
$\lvert\mu\rvert + 2\sigma$ (95%)	2.3 m	4.6 m	2.1 m	4.3 m

Table 7 Integrity Beacon flight technical error (total system error) performance on United Boeing 737 at 50 ft

95% Error, m	Vertical	Cross-track
Integrity beacon	2.1 m	4.3 m
Required navigation performance tunnel	4.5 m (TBD)	15.5 m
Meets requirement	Yes	Yes

the vertical and cross-track FTE (essentially the same as TSE for IBLS) at 50- and 100-ft altitude for the 110 autolands.

Again, using the RNP tunnel as the basis for comparison, Table 7 is constructed showing that the near-perfect NSE of IBLS allows the autopilot to reach its theoretical maximum of performance. Note that the crosswind component for approximately 20% of the landings exceeded the autopilot specification of 10 knots, sometimes by almost a factor of two. Despite this adverse condition, the system was still able to outperform any known proposed specification for TSE.

E. Flight Test Summary and Observations

The IBLS was shown to be sufficient in meeting accuracy requirements for Category III automatic landings. Additionally, the integrity of the cycle ambiguity resolution process, throughout the total of 160 test approaches (49 on the FAA King Air and 111 on the United Boeing 737), was 100% successful. No false alarms were issued, and no missed detections were registered. Note that to match the ILS sensor characteristics (which have limited dynamic range), the GPS signal had to be delayed and filtered. This extra lag is unnecessary for accuracy and suggests an avenue for autopilot improvements using the greater dynamic sensitivity of IBLS.

The consensus among all the test pilots was that the aircraft flew smooth descents with a “solid” guidance signal. FAA Administrator David R. Hinson, who flew as pilot in the left seat of the FAA King Air for approaches 46 and 47, observed that the system “seemed to be much more stable than an ILS approach”. On the King Air flights that ran on a Category III ILS runway, many of the test crew observed the “scalloping” of the ILS needles running in parallel, when compared to the GPS display, which held absolutely steady.

This benign response to both high PDOP and crosswinds, which significantly exceeded specifications, is representative of how the IBLS architecture has margin to handle stresses arising out of adverse operating conditions. In operational terms, insensitivity to high satellite PDOP translates into a significant safety margin of availability. It is equivalent to a GPS satellite failure scenario, where suddenly only a suboptimal satellite geometry is available to an aircraft on final approach. With the centimeter-level precision of IBLS ranging, PDOP less than 20 is of little consequence. That small ranging error multiplied by a geometric factor of 20 still yields a small position error.

VI. Operations Using Integrity Beacons

Depending upon the required minimums, an operational system can be supplemented with other existing sensors, such as an inertial reference unit (IRU) and a radar altimeter. The IBLS architecture is designed to satisfy operational requirements, performance specifications, and institutional issues in a way that minimizes cost and maximizes performance and efficiency.

At the airport, a pair of Integrity Beacons would be situated on either side of the approach path at approximately the range of the ILS middle marker or farther. Taking after the simple, low-cost, rugged prototype shown in Fig. 2, operational Integrity Beacons are likely to be just as routine to install and maintain as ordinary light bulbs placed around the airport. Each site can easily have redundant units broadcasting on different codes in the unlikely event of a ground failure, because Integrity Beacons do not interfere with one another.

The airborne component of operational IBLS does not differ much from that shown in Fig. 13. A benefit of the second GPS antenna (in addition to the improved level of safety on final approach) is that the aircraft has nearly 4π sr visibility of the sky and GPS. With this additional antenna, the aircraft is able to maintain lock on GPS, even during curved approaches, steep banks, and turns. The optimized airborne component also includes a loosely coupled IRU for operating at Category IIIB and IIIC minimums. For direct retrofit compatibility with existing autoland systems (either digital or analog), the GPS can emulate a traditional ILS output. The GPS receiver can also use attitude measurements from the existing IRU to calculate the lever-arm (relative position) correction between the upper and lower GPS antennas.

A. Integrity Beacon Landing System Landing Sequence

There are at least two assumed means of implementing IBLS on board aircraft: as an ILS retrofit or as a fully optimized GPS/IRU package. The ILS retrofit uses the GPS to emulate the standard ILS signal fed into the autopilot. The autopilot also makes use of the IRU measurements, but not with any coupling back to the ILS receiver. In the optimized package, GPS and IRU data are filtered together in Cartesian coordinates (not localizer and glide slope coordinates), so that the landing system has full benefit of three-axis position (and velocity) from the GPS. For an aircraft on final approach, there is not much difference between the two implementations until the conclusion of the integrity bubble pass.

The chronology of approach and landing is shown graphically in Fig. 15. Upon initiation of the approach, DGPS is used to navigate the aircraft to the integrity bubble. There—at a safe altitude—the Integrity Beacon serves as the final checkpoint before landing and defines the transition point to the high-

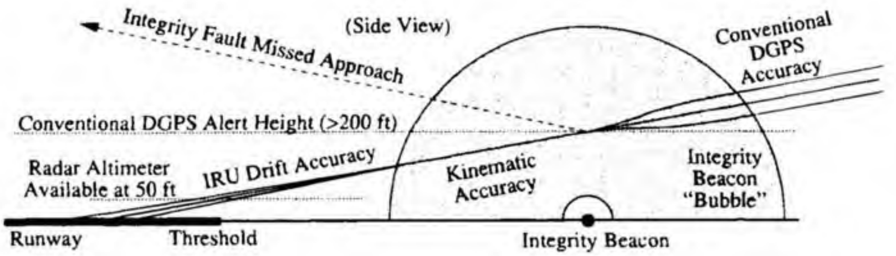

Integrity Fault Missed Approach
(Side View)
Conventional DGPS Accuracy
Conventional DGPS Alert Height (>200 ft)
Radar Altimeter Available at 50 ft
IRU Drift Accuracy
Kinematic Accuracy
Integrity Beacon "Bubble"
Runway
Threshold
Integrity Beacon

Fig. 15 Phases of Integrity Beacon system precision landing.

integrity GPS operation. Once well inside the signal bubble, the aircraft receiver has enough information to resolve cycle ambiguities and initialize its output to centimeter-level accuracy. Toward the end of the bubble pass, the GPS positioning has the highest integrity of any regime of flight. An important requirement of the design is that it preserve this level of integrity all the way through touchdown. The preferred methods for carrying this out are different for each airborne implementation of IBLS described in the following subsections.

1. *Retrofit Instrument Landing Systems*

With stand-alone GPS being used to emulate ILS (in a system topology similar to that in Fig. 13), the simplest means of preserving performance after the bubble exit is to maintain kinematic centimeter-level positioning all through touchdown. If five or more GPS satellites are visible (the same minimum requirements for enroute GPS navigation), RAIM can be continued with nearly the same effectiveness as that achieved inside the bubble. Once the cycle ambiguities are properly initialized inside the bubble, the same tight thresholds for integrity alarms applied within the bubble (as described in Sec. VII) can be applied to the carrier-phase positioning residuals. If a pseudolite is used at the airport to service the vicinity of touchdown with a modulation scheme to ensure adequate reception,[14,15] it is possible that its signal may be useable as a redundant measurement. Then, only four satellites are required for RAIM.

2. *Optimized GPS/IRU*

Given that the IRU velocity and scale factors are calibrated in flight using DGPS prior to reaching the bubble, the IRU position and velocity can be updated to kinematic accuracy at the time of the bubble pass. Thereafter, the IRU will preserve the required position accuracy (and integrity when checked against kinematic DGPS or another IRU) during the 15–20 s between the 200-ft alert height and the landing. Assuming that the system passes its internal integrity checks at the altitude of the bubble pass (at or above 200 ft), the aircraft continues its descent, navigating directly from the IRU as initialized by IBLS. Kinematic GPS continues to serve as an integrity "safety net," but even if there is any subsequent GPS failure or radio jamming of any sort, the aircraft can continue the landing, because the IRU has already been initialized. At roughly 50-ft altitude, the aircraft is over the threshold, and the radar altimeter (backed up by kinematic GPS, as demonstrated in Sec. V.A) can be employed for the flare maneuver.

Landing sequence is summarized as follows.

1) Initiate the approach using traditional differential GPS (accuracy: 2–5 m).
2) Acquire Integrity Beacon above 200-ft DGPS alert altitude.
3) Perform positioning/RAIM (integrity: missed detection probability 10^{-9}).
4) Execute go/no go decision above 200-ft alert altitude.
5) Final update of IRU position and velocity (accuracy: 2–5 cm rms).
6) Execute landing with IRU, radar altimeter, GPS; maintain 10^{-9} integrity.

VII. Integrity Beacon Landing System Navigation Integrity

The requirements on accuracy, integrity, and continuity for Category III precision approach demand the highest level of the GPS navigation performance. Specifically, combined navigation and flight control accuracy on the order of a few meters must be maintained, continuity of function preserved for all but one in 10 million (10^{-7}) approaches, and loss of integrity limited to *one in a billion* (10^{-9}) approaches.[9] Although high-accuracy navigation is possible using either differential high-performance C/A-code or differential L_1 carrier phase, the *centimeter-level* precision afforded by carrier phase provides two clear advantages. First, as previously stated, NSE represents a nearly negligible contribution to TSE. This leads to maximum margin in FTE and, therefore, maximum flexibility in flight control system design. Second, the high precision of carrier phase provides a foundation for a high level of RAIM performance.

The high performance of carrier phase can only be achieved, however, if the integer cycle ambiguities can be reliably resolved for each space vehicle (SV). IBLS is a *high-integrity* solution to real-time cycle ambiguity resolution for Category III precision approach, because the capacity for RAIM is built-in.[16] In this regard, two important observations can be made:

1) The centimeter-level precision of carrier-phase measurements provide maximum benefit from RAIM in the sense that extremely tight detection thresholds may be set without incurring unacceptably high false alarm rates. Therefore, both *high integrity* and *high continuity* are ensured.

2) The redundant ranging measurements obtained from ground-based pseudolites ensure the *availability* of RAIM.

Thus, the traditional limitations associated with high-performance navigation using RAIM-based fault detection (pseudorange measurement accuracy and low SV availability) do not exist when carrier-phase measurements are used and Integrity Beacons are present. Consequently, the integrity of IBLS-based cycle ambiguity resolution and positioning inside the IBLS bubble can be ensured through RAIM. Even after cycle ambiguity resolution, when the aircraft exits the bubble, the high precision of carrier phase is still available for kinematic positioning and RAIM, although the availability of RAIM will be somewhat degraded (to a lesser degree if a geostationary overlay is implemented). Supplementing RAIM with independent monitoring by an IRU (initialized with the high-integrity carrier-phase positioning available inside the bubble) may be beneficial in this regard and will also ensure navigation continuity even in the unlikely event of the GPS signal jamming. Ground monitoring can, of course, also be present in a supplementary role throughout the approach.

A. Receiver Autonomous Integrity Monitoring

A mathematical description of cycle ambiguity resolution is given in Sec. IV.B, and the basic RAIM theory pertinent to IBLS may be found in Ref. 16. Recall that the linearized observation equation is given by $\boldsymbol{\phi} = H\boldsymbol{u} + \boldsymbol{\delta\phi}$, where $\boldsymbol{\phi}$ is the $n \times 1$ vector of stacked single-difference phase measurements (aircraft minus reference) collected at the aircraft during pseudolite overflight. The vector $\boldsymbol{\delta\phi}$ $(\boldsymbol{n} \times \mathbf{1})$ is the single-difference phase error. The observation matrix, H $(n \times m,\ n > m)$, contains the geometric information associated with the overpass. The state vector, $\boldsymbol{u}$ ($m \times 1$), contains the cycle ambiguities and position fixes at each measurement epoch in the bubble. The position vector at an arbitrary epoch during the bubble pass $\boldsymbol{x}$ (3×1) is a vector element of $\boldsymbol{u}$:

$$\boldsymbol{u} = \begin{bmatrix} \vdots \\ \boldsymbol{x} \\ \vdots \end{bmatrix}$$

The least-squares state estimate error is as follows:

$$\delta\boldsymbol{u} = \begin{bmatrix} \vdots \\ \delta\boldsymbol{x} \\ \vdots \end{bmatrix} \approx H^{+}(\delta\boldsymbol{\phi} - \delta H\boldsymbol{u})$$

where $H^{+} = (H^{T}H)^{-1}\, H^{T}$, and δH is the error in the airborne user's knowledge of the observation matrix H. The measurement residual vector ($n \times 1$) is $\boldsymbol{r} \approx (I - HH^{+})(\delta\boldsymbol{\phi} - \delta H\boldsymbol{u})$.

Under normal conditions (NC)—no system failures—δH is negligible, and $\delta\boldsymbol{\phi}$ is normally distributed with zero mean and standard deviation of $\sigma_{\phi} = 1$ cm: $\delta\boldsymbol{\phi} = N(0, I_n\sigma_{\phi}^{2})$. Under these circumstances, it can be shown[17] that the norm of the residual vector is a χ^{2} distributed random variable with n–m DOF. A residual threshold R can then be obtained analytically to achieve any desired probability of false alarm under normal error conditions,[18] where the false alarm event is defined by the following:

$$FA \equiv (\|\boldsymbol{r}\| > \mathrm{R} \mid \mathrm{NC})$$

If a is defined to be the desired navigation system accuracy specification, then the missed detection event is given by the following:

$$MD = (\|\boldsymbol{r}\| < R,\ \|\delta\boldsymbol{x}\| > a)$$

Figure 16a is a conceptual plot of position error vs residual. The probability "ellipse" nearest the origin represents the case of such normal condition errors as multipath and receiver noise (represented by the Gaussian measurement error model given above). For a given failure mode, the ellipse will slide up the failure mode axis a distance proportional to the magnitude of the failure. In Fig. 16b a line constraint is drawn to represent the navigation system accuracy specification (a). Note that it is possible, for small failure magnitudes, that the accuracy specification will not be breached. Also shown in the figure is a threshold set on the measurement residual (R). The resulting RAIM fault detection algorithm is a simple one. Check the residual statistic to see if it is larger than the threshold. If so, a system failure is declared. Although there is complete freedom in the selection of detection thresholds, false alarms will increase as the threshold approaches zero. However, the detection threshold can always be chosen to produce a low false alarm rate under normal error conditions. In the case of GPS carrier phase measurements, the overall result is shown conceptually in Fig. 16c. A hypothetical failure mode penetrating the narrow missed detection region is mathematically possible to construct, but given that such a mode must be related physically to a *real system failure,* the likelihood of its occurrence will be low.

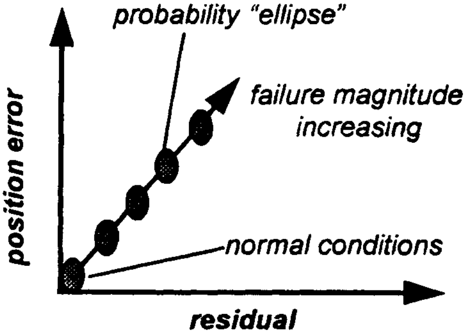


Fig. 16a State estimate vs residual.

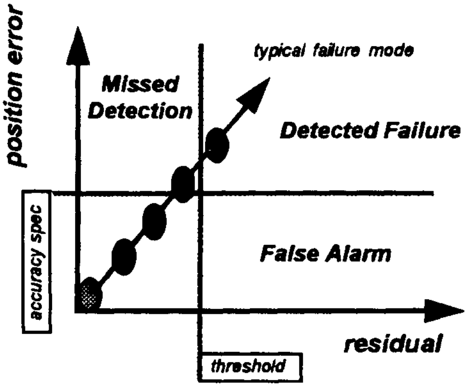


Fig. 16b Basic receiver autonomous integrity monitoring.

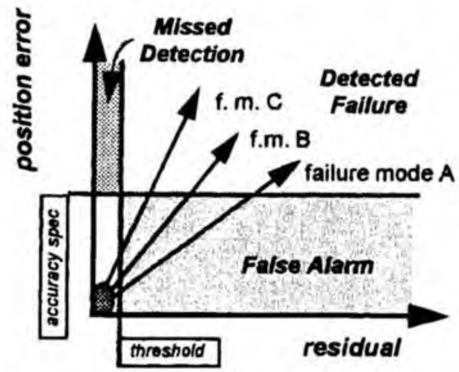


Fig. 16c High-performance receiver autonomous integrity monitoring with Integrity Beacon Landing System.

Clearly, the *quantitative* verification of RAIM fault detection capability depends upon the nature and likelihood of navigation system fault modes.

B. System Failure Modes

A loss of integrity event occurs when a navigation system failure, or an unusually large ranging error attributable to a familiar source such as multipath, causes a large position error that is undetectable by any form of monitoring. A top-level fault tree illustrating a number of integrity-threat failure classes associated with the space, airborne, and ground segments of IBLS is shown in Fig. 17. For this example, three diverse failure modes are chosen, one from each of the three segments of IBLS.

1. Airborne Segment: Cycle Slips

Cycle slips are most often associated with the airborne receiver; however, once differencing (or differential correction) is done, a cycle slip in the reference receiver will result in the same overall ranging error as a cycle slip in the aircraft receiver and must, therefore, be detected. The probability of cycle slip occurrence is dependent upon the particular receiver and antenna used. In general, however, the probability of a cycle slip event increases with increasing phase–locked-loop bandwidth, increasing time, and decreasing signal strength. As a first layer of the safety net, the low signal strength conditions under which cycle slips are likely to occur are identified at the signal processing (phase–locked-loop) level. Thereafter, RAIM provides an important additional layer of cycle slip monitoring.

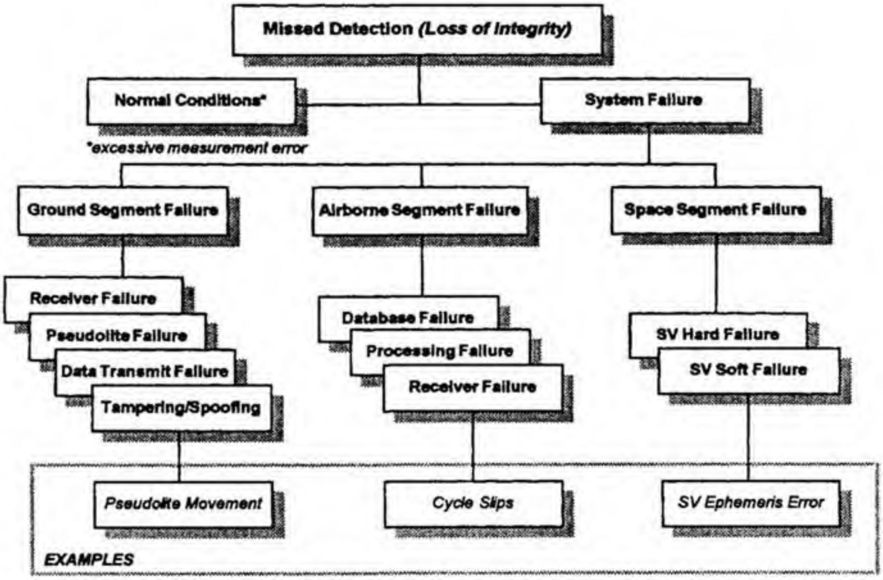


Fig. 17 Navigation system fault tree.

2. *Ground Segment: Movement of Ground Hardware*

This failure mode may be expressed as a mismatch between the actual ground hardware location and the location given in the airborne database. The question has been raised that perhaps a differential-based landing system such as IBLS, because of the existence of ground hardware, may be susceptible to either intentional tampering or unintentional errors in antenna siting (perhaps the result of maintenance work). Such discrepancies can also result from errors in the airborne IBLS ground survey database or, possibly, as errors in flight inspection. Movement of the reference station to an unsurveyed location before the approach begins, while leaving the pseudolites untouched, will not affect cycle ambiguity resolution or centimeter-level positioning.[19] However, movement larger than the code measurement noise, will generally trigger an alarm caused by the discrepancy between code and carrier. Furthermore, RAIM ensures that IBLS is also robust to movement of Integrity Beacon pseudolites.

3. *Space Segment: Spacecraft Ephemeris Errors*

Spacecraft soft failures—those spacecraft failures unknown to the user—fall into two basic categories: SV clock errors and SV ephemeris errors. Although the effects of clock errors originating at the spacecraft are almost completely eliminated through differential positioning, those of spacecraft ephemeris errors are not. Rather large ephemeris errors (> 500 m) are required to produce noticeable positioning errors. Among the possible origins of such an error are intentionally induced SA errors of unusually large magnitude, orbit determination error, and errors in ephemeris upload. The resulting user position estimate error, and RAIM measurement residual will both scale linearly with the displacement between the aircraft and the reference station.

An exception to the differential cancellation of ranging errors was recently exhibited by SV 19.[10] The symptom was a pseudorange bias of up to several meters when nonidentical receivers were used at the reference station and aircraft. Note that a code-ranging error of this type may or may not be detectable by ground monitoring, depending upon the actual receivers used at the reference station, monitor station, and aircraft. It is noteworthy, however, that IBLS tests during the occurrence of this ranging anomaly *demonstrated that IBLS carrier-phase tracking of SV 19 was not affected.*

C. Quantifying Integrity

The statistical significance necessary to demonstrate integrity $P(MD)$ on the order of 10^{-9} cannot, of course, be attained through flight test or other experimental means. The large total number of sample approaches and wide range of system failures can only be achieved through mathematical models and computer simulation. A valid method for quantifying integrity through simulation is discussed in Ref. 19 and can be applied both to normal error conditions and the three representative types of system failure already chosen. Considering the case of normal system errors (receiver noise and multipath) first, Fig. 18 shows a plot of $P(\|\delta x\| > a)$—a conservative measure of integrity under normal error conditions. For comparison, the equivalent result for high-precision C/A-code ranging

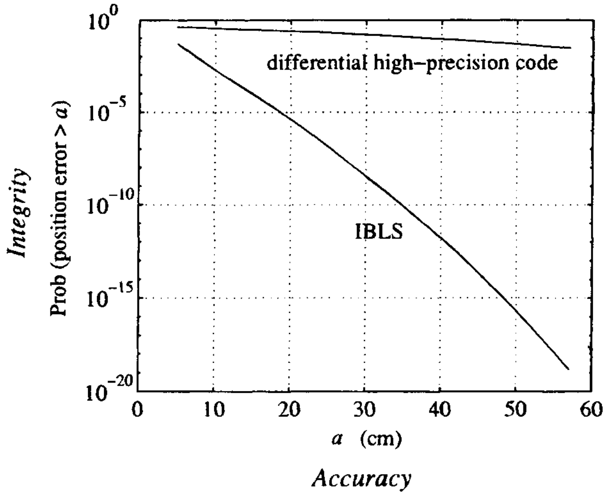


Fig. 18 Integrity under normal error conditions.

is also included. The level of navigation system integrity under normal error conditions is better than 10^{-10} even for accuracies of 35 cm with IBLS; whereas, for the equivalent accuracy the integrity of a system based on high precision code is roughly only 10^{-1}.

In addition, the effectiveness of the RAIM-based fault detection capability built into IBLS (Sec. VII.A) is demonstrated by the results of over 25 million simulated approaches using representative models for the three fault modes considered above.[20] Figure 19 shows the surface relating system integrity [$\log_{10}P(MD)$], continuity [$\log_{10}P(FA)$], and accuracy (a). As intuitively expected,

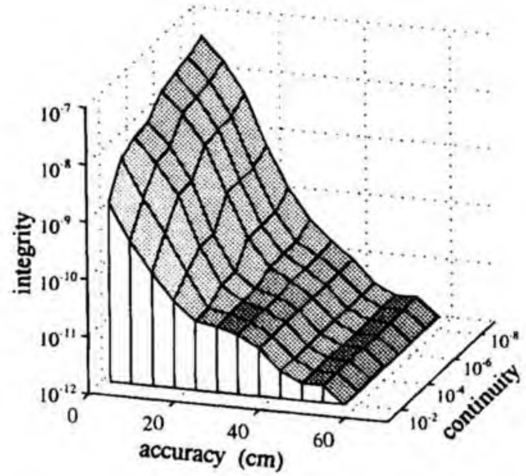


Fig. 19 Navigation system integrity, accuracy, and continuity.

integrity improves as the requirements on navigation system accuracy and continuity are relaxed. The resulting surface can be interpreted in an absolute sense as well. Integrity in the IBLS bubble to the three failures simulated is better than 10^{-10}, even for an accuracy requirement of 35 cm and continuity (false alarm probability under normal error conditions) of better than 10^{-7}. In summary, in the presence of normal error conditions and the three failure modes considered, Category III levels of integrity and continuity are achievable with submeter navigation system accuracy.

D. Signal Interference

This subsection addresses the issue of interference to the radio signals used on final approach and landing. The section deals with the potential effects of hostile jammers and spoofers, and it discusses the impact of the so-called "near–far" problem with respect to the Integrity Beacon pseudolites. As with any radionavigation system, spoofing (intentional hostile generation of a false radio signal in an effort to mislead an aircraft into an unsafe condition) is a more important issue than jamming.

1. Jamming

Fortunately, jamming is a readily detectable condition (either through receiver loss of lock, or by monitoring.cross-correlation or AGC levels). Therefore, jamming is generally only a nuisance rather than a life-threatening situation. In a hostile jamming campaign scenario, pilots and ground controllers have a number of options with which to respond, including diverting the aircraft to another airport. Using the same kinematic techniques used for landing, equipment could be developed to locate a hostile jammer rapidly. To assess the likelihood of the jamming scenario, it is useful to consider that, although portable, hand-held aviation-band VHF radios are readily available to the general public for just a few hundred dollars, the instances where air-to-tower voice communications have been intentionally obstructed are extremely rare.

2. Spoofing

The GPS Integrity Beacon concept provides a significant barrier to intentional, hostile tampering with the ground system. Although the authors are unaware of any reported instances of hostile tampering with the ILS equipment, this potential vulnerability of any radionavigation aid must be considered. Unlike conventional DGPS, the built-in redundancy of IBLS signals makes it nearly impossible to corrupt the signal with anything that spoofs the aircraft into thinking it is anywhere other than where it really is.

An example of this robustness against spoofing is shown in Fig. 20. In the scenario shown, the spoofer attempts to direct the aircraft into the ground by broadcasting false GPS differential corrections. These corrections assume a lower altitude for the reference station. Conventional RAIM algorithms would not be able to detect this condition, because the residuals would all be self-consistent. With the Integrity Beacon, however, the aircraft immediately registers that the pseudolite phase signature is completely inconsistent with its conventional posi-

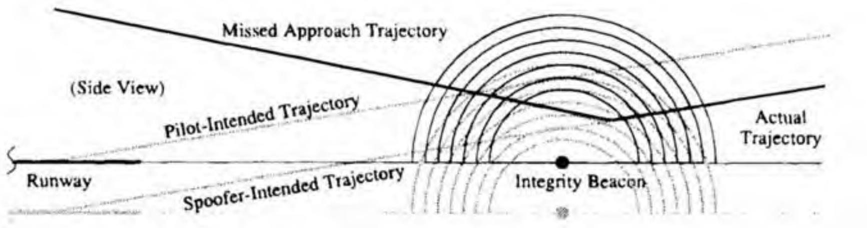

Missed Approach Trajectory
(Side View)
Pilot-Intended Trajectory
Actual Trajectory
Runway
Spoofer-Intended Trajectory
Integrity Beacon

Fig. 20 Integrity Beacon spherical wavefronts make spoofing extremely difficult.

tion fixes. The system issues an alarm to the pilot, who executes a missed approach.

Spoofing scenarios in which simulated GPS Integrity Beacon signals are broadcast are extremely difficult to execute, given that the spoofer must know the aircraft position and attitude to high accuracy in order to calculate what signal to transmit. The key point is that such spoofing would be so difficult that any would-be tamperer will most likely direct his efforts elsewhere.

3. *Near–Far Problem*

One perceived shortcoming that is frequently pointed out in applications employing pseudolites is the so-called near–far problem. For the special case of the *low-power* Integrity Beacon, the architecture has been intentionally designed so that this phenomenon is of little consequence. The cross-correlation (isolation) between the GPS gold codes in two separate 1023-bit PRN sequences may be no better than 21.6 dB worst case.[21] The power level of −130 dBm received from GPS satellites is relatively constant, because a terrestrial user's range to the satellite does not vary much as a percentage of the average range. The problem is that when a user is operating at close range to pseudolites, there is a possibility that the pseudolite power may exceed the ambient GPS power by the cross-correlation threshold.

Figure 21 shows a side view of an Integrity Beacon and how the bubble geometry readily sidesteps this issue. The effective dynamic range of the signal is indicated by the shaded region. Based on the nominal glide slope shown, it is clear that there is significant margin for the approach altitude to be very much higher or lower than nominal while still staying within the correct power level. If the aircraft receiver registers a power level that becomes much too high, it is a clear indication that the aircraft is flying too close to the ground and should execute a missed approach.

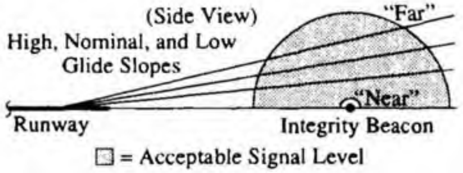

(Side View)
High, Nominal, and Low Glide Slopes
"Far"
"Near"
Runway
Integrity Beacon
▨ = Acceptable Signal Level

Fig. 21 Integrity Beacon Landing System sidesteps the near–far problem.

In future implementations, some thought is being directed toward increasing the bubble size for a larger coverage volume and perhaps extending the boundary to encompass the touchdown zone. Another concept is to augment the Integrity Beacons with a pseudolite back at the airport to provide redundant carrier ranging and a built-in data link (as mentioned in Sec. VI.A). Of course, these enhanced implementations would be subject to the near–far problem (as well as antenna pattern and placement considerations) and, therefore, might employ pulsing or other modulation schemes[14,15] to reduce interference with the GPS signals.

VIII. Conclusion

Analysis and flight testing are both demonstrating that IBLS is capable of meeting and exceeding the stringent requirements of Category III precision landing, especially in the most challenging regimes of accuracy and integrity. The centimeter-level precision of GPS carrier phase and the built-in crosschecks and redundancy provided by Integrity Beacons translate into robust, dependable performance and continued assurance to the pilot that it is safe to land. The end result is an architecture that is safe for passengers and crew.

References

[1] Cohen, C. E., Pervan, B. S., Cobb, H. S., Lawrence, D. G., Powell, J. D., and Parkinson, B. W., "Real-Time Cycle Ambiguity Resolution using a Pseudolite for Precision Landing of Aircraft with GPS," DSNS '93, Amsterdam, The Netherlands, March 30–April 2, 1993.

[2] Cohen, C. E., Pervan, B. S., Lawrence, D. G., Cobb, H. S., Powell, J. D., and Parkinson, B. W., "Real-Time Flight Test Evaluation of the GPS Marker Beacon Concept for Category III Kinematic GPS Precision Landing," *Proceedings of ION GPS-93,* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 22–24, 1993.

[3] Cohen, C. E., Pervan, B. S., Cobb, H. S., Lawrence, D. G., Powell, J. D., and Parkinson, B. W., "Achieving Required Navigation Performance using GNSS for Category III Precision Landing," DSNS-94, London, UK, April, 1994.

[4] Hatch, R., "Instantaneous Ambiguity Resolution," KIS Symposium 1990, Banff, Canada, Sept. 1990.

[5] Lachapelle, G., Cannon, M. E., and Lu, G., "Ambiguity Resolution on-the-Fly: A Comparison of P-Code and High Performance C/A-Code Receiver Technologies," *Proceedings of ION GPS-92,* (Albuquerque, NM), Institute of Navigation, Washington, DC, Sept. 16–18, 1992.

[6] Cohen, C. E., "Attitude Determination Using GPS," Ph.D. Dissertation, Stanford University, Stanford, CA Dec., 1992.

[7] Anon., "*International Standards, Recommended Practices and Procedures for Air Navigation Services—Annex 10,*" International Civic Aviation Organization, April, 1985.

[8] "U.S. Federal Radionavigation Plan," U.S. Departments of Transportation and Defense, 1992.

[9] Kelley, R. J., and Davis, J. M., "Required Navigation Performance (RNP) for Precision Approach and Landing with GNSS Application," *Navigation,* Vol. 41, No. 1, 1994.

[10] Nordwall, B. D., "Filter Center Column," *Aviation Week and Space Technology,* Aug. 30, 1993.

[11]Cohen, C. E., Lawrence, D. G., Pervan, B. S., Cobb, H. S., Barrows, A. K., Powell, J. D., Parkinson, B. W., Wullschleger, V., and Kalinowski, S., "Flight Test Results of Autocoupled Approaches Using GPS and Integrity Beacons," *Proceedings of ION GPS-94,* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[12]Paielli, R., Bach, R., and McNally, D., work in progress, initially reported as "Carrier Phase Differential GPS for Approach and Landing: DGPS/INS Integration and Flight Test Validation," ION National Technical Meeting, Anaheim, CA, Jan. 1995.

[13]Cohen, C. E., Lawrence, D. G., Cobb, H. S., Pervan, B. S., Powell, J. D., Parkinson, B. W., Aubrey, G. J., Loewe, W., Ormiston, D., McNally, B. D., Kaufmann, D. N., Wullschleger, V., and Swider, R., "Preliminary Results of Category III Precision Landing with 110 Automatic Landings of a United Boeing 737 using GNSS Integrity Beacons," ION National Technical Meeting, Anaheim, CA, Jan. 1995.

[14]Stansell, T. A., "RTCM SC-104 Recommended Pseudolite Signal Specification," *Navigation,* Vol. III, 1986.

[15]Elrod, B., Barltrop, K., and Dierondorck, A. J., Van "Testing of GPS Augmented with Pseudolites for Precision Approach Applications," *Proceedings of ION GPS-94,* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[16]Pervan, B. S., Cohen, C. E., and Parkinson, B. W., "Integrity Monitoring for Precision Approach Using Kinematic GPS and a Ground-Based Pseudolite," *Navigation,* Vol. 41, No. 2, 1994.

[17]Parkinson, B. W., and Axelrad, P., "Autonomous GPS Integrity Monitoring Using the Pseudorange Residual," *Navigation,* Vol. 35, No. 2, 1988.

[18]Sturza, M. A., "Navigation System Integrity Monitoring Using Redundant Measurements," *Navigation,* Vol. 35, No. 4, 1988–89.

[19]Pervan, B. S., Cohen, C. E., and Parkinson, B. W., "Integrity in Cycle Ambiguity Resolution for GPS-Based Precision Landing," DSNS '94, London, UK, April 18–22, 1994.

[20]Pervan, B. S., Cohen, C. E., Lawrence, D. G., Cobb, H. S., Powell, J. D., and Parkinson, B. W., "Autonomous Integrity Monitoring for Precision Landing Using Ground-Based Integrity Beacon Pseudolites," *Proceeding of ION GPS-94,* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[21]Spilker, J. J., Jr., "GPS Signal Structure and Performance Characteristics," *Navigation,* Vol. I, 1980, pp. 29–54.

Chapter 16

Spacecraft Attitude Control Using GPS Carrier Phase

E. Glenn Lightsey*
NASA Goddard Space Flight Center, Greenbelt, Maryland 20771

I. Introduction

VIRTUALLY all but the simplest spacecraft employ some means of active attitude control using such actuators as control momentum gyros, reaction wheels, offset thrusters, and magnetic torque rods. Attitude control is almost always performed by a closed-loop system onboard the spacecraft; especially for low Earth orbit (LEO) spacecraft, where ground contacts are often limited to a few minutes per day. Only unusual events such as a momentum unload or spacecraft slew maneuver are commanded from the ground, and even then, the onboard system is usually responsible for some level of automatic control. Closed-loop attitude control, of course, requires sensor feedback of the vehicle orientation. This has traditionally been provided by such low-cost sensors as magnetometers, horizon sensors, and sun sensors, or more expensive high-performance instruments including gyroscopes and star trackers. Recently developed GPS attitude determination systems provide an opportunity to use this new technology in attitude control system designs.

As described in Chapter 19, this volume, precise measurement of differential carrier phase between multiple antennas may be used to determine the attitude of a vehicle. Thus, installing a GPS receiver onboard a spacecraft affords the opportunity to use a single lightweight, low-cost sensor for a multitude of functions: position, velocity, attitude, attitude rate, and time. This consolidation of resources is likely to lead to an overall savings in cost, power, weight, and complexity for spacecraft. Furthermore, the elimination of many different sensor devices and their interfaces can yield a substantial benefit in system reliability.

The key performance issues regarding the utility of GPS as a closed-loop attitude sensor are bandwidth, accuracy, and antenna placement. Typical low-

*Engineer, Guidance and Control Branch; also Ph.D. Candidate, Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305.

precision attitude control systems using such sensors as magnetometers, Earth sensors, and Sun presence detectors, have bandwidths on the order of a few times orbit rate and pointing accuracies of 1–5 deg (example: the Solar Anomalous and Magnetospheric Particle Explorer, SAMPEX). For high-performance applications using gyros and star trackers, bandwidths range up to a few Hz with pointing accuracies of 0.1 deg or better (example: the Hubble Space Telescope, HST). The update rate of the currently designed GPS attitude determination system is in the range of 0.1–10 Hz, although the theoretical defined limit identified by Cohen[1] is 1 kHz. The accuracy of GPS attitude determination is strongly dependent on antenna placement and data-processing techniques. For a one meter antenna separation, the point solution accuracy is approximately 0.3 deg, with possible improvements to better than 0.1 deg using dynamic filtering. Both accuracy and bandwidth performance can be extended by combining the GPS sensor with gyroscopes.

The operational capability of the GPS attitude sensor was demonstrated in space in June 1993 when the Air Force sponsored RADCAL satellite was launched into an 800-km polar orbit. This gravity-gradient stabilized satellite contained two cross-strapped GPS receivers from which differential carrier phase measurements were used to obtain attitude solutions in postprocessing (see Fig. 1 and Refs. 2 and 3). One receiver failed after six months in orbit, but the two receivers together have provided more than 18 months of measurement data since launch.

Further plans are in progress to use a GPS receiver as a real-time attitude sensor on a host of other spacecraft including the OAST Flyer (discussed below), Gemstar, REX-II, Orbcomm, Globalstar, SSTI Lewis, SSTI Clark, and others. If the attitude determination performance using GPS is found to be acceptable in terms of both accuracy and bandwidth, a wide variety of control schemes can be used based on the type of actuators available. This development is likely to have tremendous benefit for LEO spacecraft missions, including remote sensing and mobile communications. The remainder of this chapter provides further details on GPS attitude sensing for spacecraft and a specific case study of the design of a control system using GPS as the attitude sensor.

II. Design Case Study

One of the first demonstrations of closed-loop attitude control in space using GPS will be the GADACS experiment (GPS Attitude Determination and Control System, pronounced "gay-dax"), which is manifested for flight onboard Space Shuttle mission STS-69 in Autumn 1995. The satellite bus is the rectangular SPARTAN "OAST Flyer" payload, measuring approximately 1 × 1.25 × 1.5 m, as shown in Fig. 2. It is deployed from the Space Shuttle cargo bay as a free-flying spacecraft for approximately 40 h of operation before retrieval. The vehicle orientation is controlled in three axes through the actuation of small cold nitrogen gas jets. The OAST Flyer mission will be carrying at least two additional experiments, the Return Flux Experiment (REFLEX) and the Solar Exposure to Laser Ordnance Devices Experiment (SELODE).

The REFLEX and SELODE experiments will be conducted for approximately the first two-thirds of the 2-day mission. During this time, the SPARTAN attitude

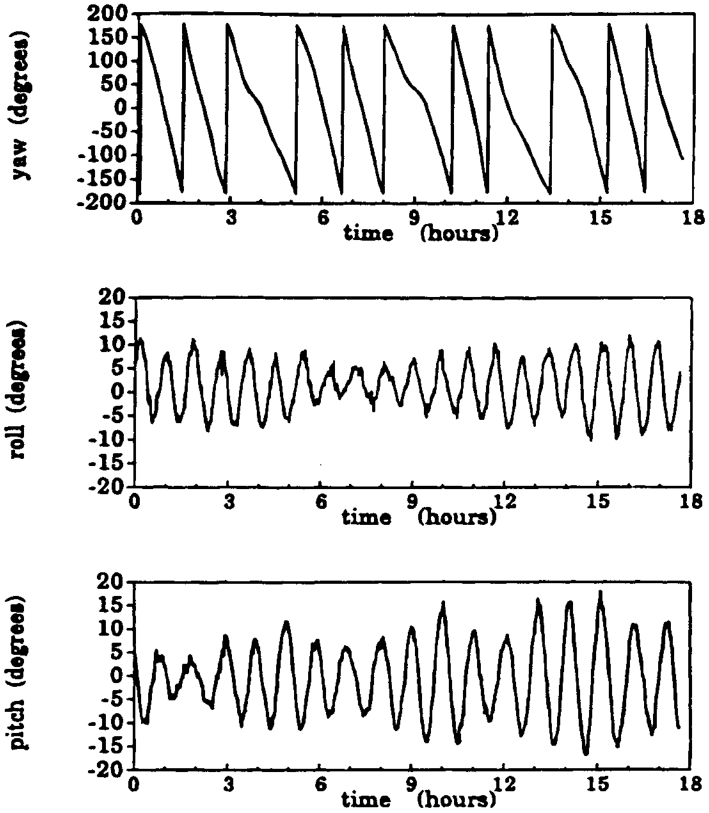


Fig. 1 GPS carrier-phase-based attitude flight results (RADCAL data, unfiltered).

control system will maneuver the spacecraft into a preset series of pointing profiles while these experiments are operating.

The attitude determination part of the GADACS experiment will be operational the entire time the OAST Flyer is free flying. During the REFLEX and SELODE portions of the mission, GADACS will be collecting attitude determination data using its two GPS attitude receivers. Control of the spacecraft will be maintained by the nominal OAST Flyer controller. GPS data will be recorded on tape along with gyro measurements that will be used to verify and calibrate the real-time GPS attitude measurements. Star tracker updates will be performed every other orbit throughout the mission to calibrate the onboard gyros.

For the last portion of the mission, GADACS will assume control of the vehicle, using only GPS-sensed attitude for closed-loop attitude control. During this time, GADACS will control the spacecraft in a series of Earth-pointing and inertial profiles and execute a series of test slews in order to test the performance of the GPS attitude-based control for a variety of mission types. The changeover from one receiver/antenna set to another will also be exercised by these maneu-

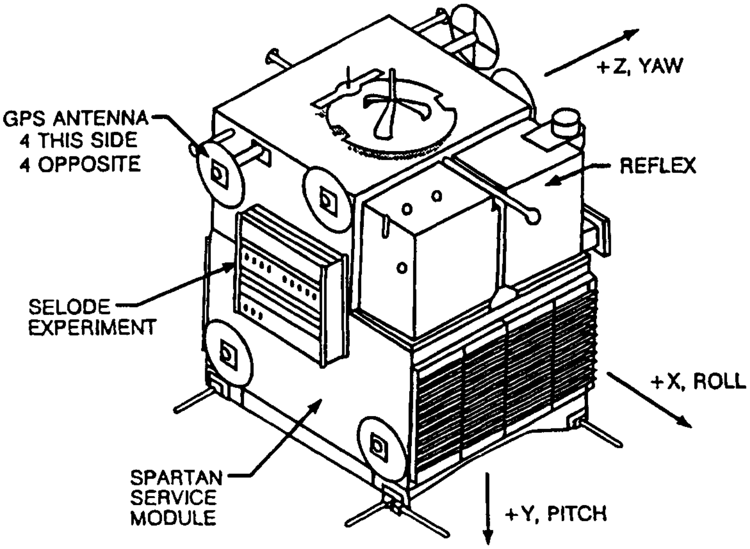


Fig. 2 SPARTAN OAST-flyer spacecraft.

vers. A detailed discussion of the GADACS mission and hardware is given in Ref. 4.

The three nonlinear single-axis control systems that GADACS will employ were based on the GPS sensor behavior, the expected environmental disturbance torques, and the performance and mission requirements imposed on the experiment by the spacecraft. Of the latter, the dominant consideration was conservation of an extremely limited actuation fuel budget. The final design is presented as a case study along with a discussion of the issues leading to its selection and expected performance.

III. Sensor Characteristics

The fundamental performance characteristics of any sensor should be understood before attempting to use it in a control application. In the case of a GPS attitude receiver, the main determinants of system performance are antenna separation (also known as baseline length) and noncommon mode error sources; the dominant component of the latter being multipath interference.

Because relative position error is roughly independent of antenna separation, greater baseline length leads to more precise attitude measurements. For a given antenna separation, as shown in Fig. 3, the baseline vector between the master antenna position $\boldsymbol{a}_o$ and the slave antenna position $\boldsymbol{a}_i$ is defined as follows:

$$\boldsymbol{b}_i = \boldsymbol{a}_i - \boldsymbol{a}_o \tag{1}$$

This baseline vector is assumed to be known in the vehicle body fixed frame B.

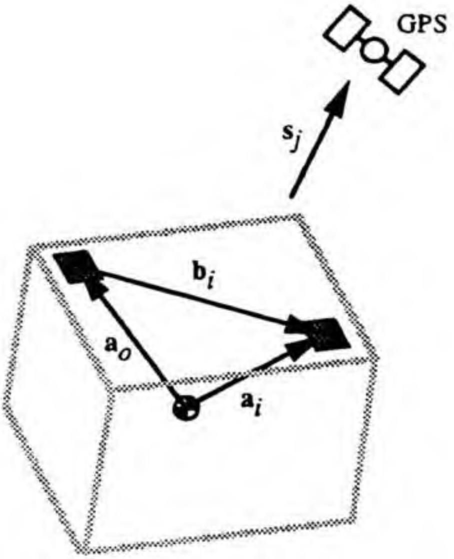


Fig. 3 Attitude observability geometry.

The carrier phase difference measured by the GPS receiver provides a measure of the projection of the baseline vector $\boldsymbol{b}_i$ onto the line of sight from the vehicle to the GPS satellite $\boldsymbol{s}_j$:

$$\Delta\varphi_{ij} = \boldsymbol{b}_i \cdot \boldsymbol{s}_j \tag{2}$$

where the integer cycle ambiguity has been neglected. This subject is discussed in detail in Chapter 19, this volume.

Because the line-of-sight vector is known in the orbit local frame, and the baseline vector is known as the body frame, the phase difference can be expressed as a function of the attitude of the vehicle:

$$\Delta\varphi_{ij} = (\boldsymbol{b}_i^B)^T A s_j^L \tag{3}$$

where the superscripts indicate the coordinate frame in which the vector is expressed. A is the attitude matrix of the vehicle, or in other words, the rotation matrix from the orbit local to the body fixed frame.

Equation (3) can be linearized about a nominal or previously estimate attitude as follows:

$$\Delta\varphi_{ij} = (\boldsymbol{b}_i^B)^T(\delta A A_o)\boldsymbol{s}_j^L \tag{4}$$

In this case, A_o is the nominal attitude matrix. Chapter 14, this volume, provides details on how measurement equations from several satellites and baselines can be solved for the best attitude estimate.

This relation shows that for given $\Delta\varphi$, A_o, and $\boldsymbol{s}$, there is a direct relationship between δA and $\boldsymbol{b}$. To achieve smaller resolution in attitude δA a larger baseline vector $\boldsymbol{b}$ is needed, as expected.

A. Antenna Placement

For most satellites, the payload size is required to be minimized. The maximum antenna separation should, therefore, be sought to achieve the best possible attitude resolution. This usually results in antenna placement on the corners of the zenith facing side of the satellite, if possible. Even for the relatively small baselines of many spacecraft (often <1 m), subdegree accuracy is still possible using GPS.

If the satellite is not Earth-pointed (for example, an inertially pointed platform), then potential blockage of GPS signals by the Earth must be considered. If continuous attitude information is required in this case, it may be necessary to have two independent GPS receivers with antennas installed on opposite faces of the spacecraft. At least three antennas (preferably four or more) must share the same sky view to make the differential carrier phase measurements necessary for attitude determination.

Another antenna placement issue is the location of the antenna array relative to the key control requirements. For example, if the antennas are mounted on the solar panels, and the pointing requirements are driven by an instrument mounted on the main body of the spacecraft, there will be a loss in accuracy. Furthermore, antennas mounted on flexible or moving parts may reduce accuracy. A design tradeoff exists between geometric resolution (increasing antenna separation) and sensor collocation with the point of interest (decreasing antenna separation). For most small spacecraft, such as GADACS, the additional geometric resolution gained by maximizing the antenna separation on nonmoving spacecraft parts results in an overall improvement in sensor accuracy.

B. Sensor Calibration

The theoretically achievable attitude accuracy is limited by mechanical system knowledge. Mechanical system knowledge is defined as the precision to which the antenna phase center locations are known on the spacecraft and the alignment of the spacecraft body axes to the antenna axes. These are important quantities that should be determined as accurately as possible during ground testing. Many GPS receivers measure the antenna locations autonomously by placing the receiver into a self-survey mode and allowing the GPS satellites to pass overhead for several hours. This test requires the spacecraft truss to have a sky view, and it should be planned in advance. The axis alignment usually must be measured by some independent means (a theodolite, for example).

For the GADACS experiment, a significant effort was undertaken to measure the antenna baseline lengths and the axis alignment precisely. Five 7-hour self-survey tests were performed over a 4-day period using the spacecraft truss in an open area selected to minimize multipath reflections. The antenna baselines were found to be repeatable to within 2.2 mm (roughly 1 part in 500). During this time, the antenna axis alignment was validated optically using a theodolite and an antenna-mounted optical reflection cube to within 1 arc minute. A detailed description of the calibration test and results is given in Ref. 5.

In the future, it is expected that an alternative approach to an outdoor test on the spacecraft truss will be developed. Some methods that have been considered to determine the same information are indoor calibration using simulated GPS signals,[6] and on-orbit sensor calibration performed in postprocessing[7] or near real time.[8]

C. Multipath

Another possible error source is multipath. Multipath occurs when the signal arrives at the antenna from reflected surfaces in addition to the line-of-sight source. The reflected signal is phase shifted with respect to the original transmission and appears as additive noise at the antenna. Because the antenna locations are different, the multipath signature at each antenna is unique and the error is not common mode.

Multipath is the dominant error source in many spacecraft applications, accounting for more than 90% of the total error budget in carrier phase measurement. Unfortunately, its presence is pervasive, although steps can be taken to minimize it. Isolating the hemispherical patch antennas, providing unobstructed fields of view, and adding ground planes will all reduce multipath. Canting the antennas away from reflective surfaces may also improve the measurement quality, although this technique also reduces the field of view common to all antennas. The options are usually limited on a typical mission with many conflicting design constraints. Empirical tests have shown that for complex reflective surfaces, such as spacecraft, a conservative value for carrier phase error caused by multipath is approximately 5 mm rms (see Refs. 1 and 9).

D. Sensor Accuracy

These and other error sources are discussed more rigorously in other works.[9,10] If the mechanical system is well known, the total system performance is characterized by the multipath environment. The relationship between rms attitude error, range error (carrier phase), and baseline length is approximately as follows:

$$\sigma_\theta = \sigma_r / b \tag{5}$$

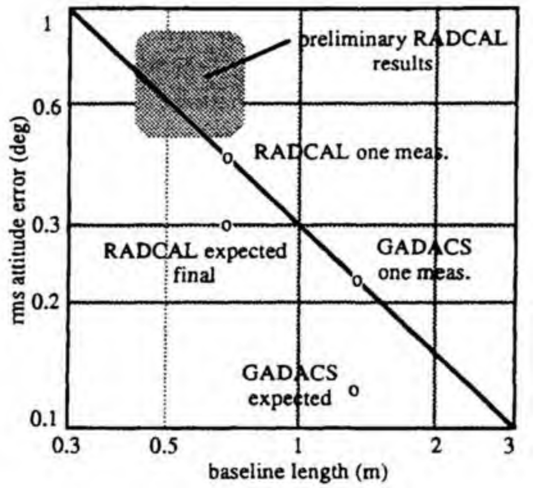
A plot of this empirically verified relationship along with datapoints for RADCAL and GADACS baselines is shown in Fig. 4. From this chart, about 0.5 deg attitude accuracy could be expected for a single RADCAL measurement. If the three simultaneous measurements are assumed to be statistically independent (not strictly true, but a useful approximation), a single, more accurate estimate is produced with the following covariance:

$$1/\sigma_m^2 = 1/\sigma_1^2 + 1/\sigma_2^2 + 1/\sigma_3^2 = 9.1 \text{ deg}^{-2} \tag{6}$$

or

$$\sigma_m \sim 0.3 \text{ deg} \tag{7}$$

Preliminary flight results from RADCAL (Fig. 1) indicate that the accuracy of the attitude solution is in the shaded region shown on Fig. 4 with σ_m approximately 0.5–1.0 deg. The reason for the difference from the theoretical performance is that the mechanical system is not well known. Accuracy better than 0.5 deg rms is expected when the data reduction is finished. However, because there is no other independent attitude sensor on RADCAL (other than a magnetometer,



baseline length (m)

Fig. 4 Attitude determination accuracy.

which does not provide better than degree accuracy), the actual GPS attitude sensor accuracy is not precisely known.

The GADACS experiment, which has antenna baselines that are approximately twice as long as RADCAL (1.20 m vs 0.67 m), can expect $\sigma_m \sim 0.2$ deg. This is an important design parameter that will be used to size the thruster dead band in the attitude control loop.

E. Dynamic Filtering

Improved attitude estimates can be derived by using a Kalman filter to include knowledge of the vehicle dynamics in the estimation process. Fig. 5 illustrates the results for the same dataset shown in Fig. 1 with an extended Kalman filter. The filter states consist of three elements of a correction quaternion, three angular velocities, and three line biases. If the highest accuracy estimates are required for the controller, it is probably best to use a filter. However, if the dynamic model is not correct, the filter will produce unreliable and possibly divergent results.

IV. Vehicle Dynamics

The control system design begins with a study of the uncontrolled vehicle dynamics. To size the body-derived environmental torques, it is convenient to use an "orbit" reference frame to express vehicle attitude, as shown in Fig. 6. The frame is formed by taking the geodetic nadir as the z_o axis, crossing it into the inertial velocity vector to form the y_o axis, and completing the orthogonal set by computing x_o as the cross product of y_o and z_o. The x_o axis is then aligned with the local horizontal component of the velocity vector, which approximately coincides with the velocity vector for near-circular orbits.

The vehicle body axes (shown in Figs. 2 and 6) are assumed to be aligned with the principal axes of inertia. In fact, they are misaligned by about 10 deg,

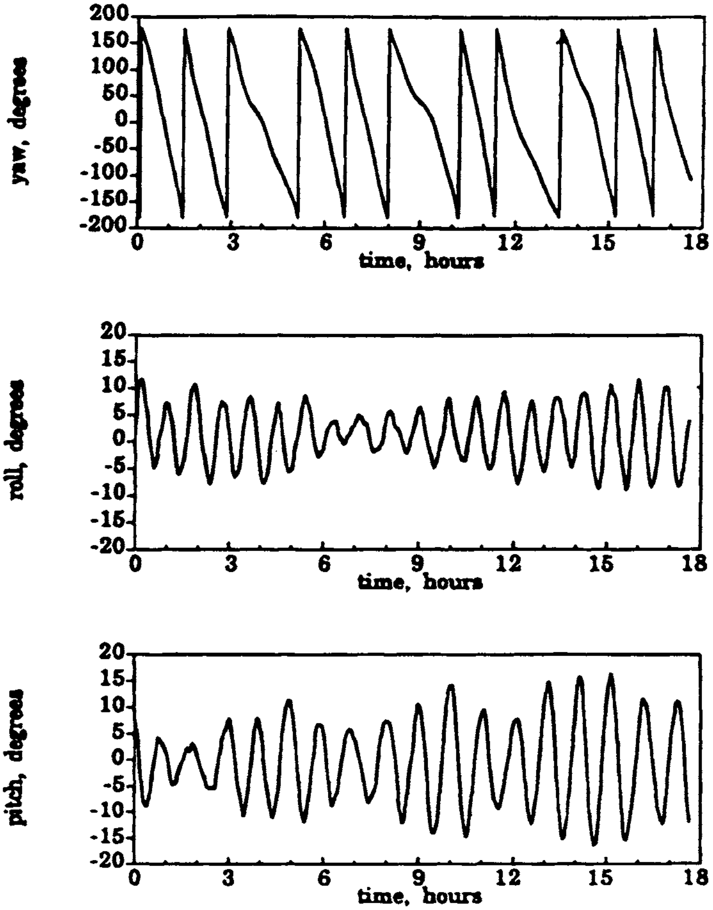


Fig. 5 GPS carrier-phase-based attitude flight results (RADCAL data, with Kalman filter).

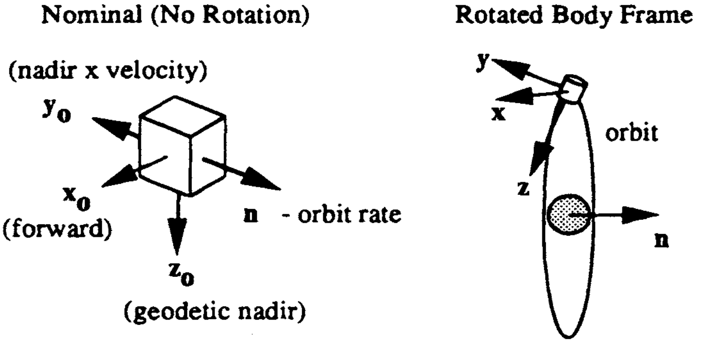


Fig. 6 Attitude reference frames.

but that effect is neglected in this simplified analysis. The vehicle attitude is then expressed as a 321 (yaw, pitch, roll) Euler rotation sequence from the orbit reference frame.

With the vehicle frame defined, the next step is to model the significant disturbance torques in terms of the vehicle attitude. At the Space Shuttle altitude of 380 km, aerodynamic drag is usually the dominant environmental torque. For small, relatively symmetric satellites, however, where the deviation between the center of pressure and center of mass is not large, the gravity gradient and drag torques may be equally important. In fact, the peak gravity gradient torque may be determined from the SPARTAN inertia listed in Table 1 and compared to the previously observed maximum aerodynamic torque for the Spartan payload[11]:

$$\max|M_{gg}| = 5.77e{-5}\ \mathrm{N\ m} \tag{8}$$

$$\max|M_a| = 3.39e{-4}\ \mathrm{N\ m} \tag{9}$$

The aerodynamic torque is seen to be about 5.9 times greater than the gravity gradient torque. Because this is less than an order of magnitude, a thorough analysis of the vehicle dynamics should include both effects.

A. Gravity Gradient Moment

The gravity gradient torque is examined first. Other references (e.g., Ref. 12) have demonstrated that for the 321 Euler sequence of a vehicle in a near-circular orbit whose body axes coincide with the principal axes of inertia (all true for GADACS), the gravity gradient torque may be linearized for small ψ (yaw), θ (pitch), and ϕ (roll) as:

$$\boldsymbol{M}_{gg} = -3n^2[(J_y - J_z)\phi\mathbf{i} + (J_x - J_z)\theta\boldsymbol{j}] \tag{10}$$

where the $[\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}]$ unit vector set defines the $[x, y, z]$ vehicle reference frame. The variable n is orbit rate.

A useful parameterization first performed by DeBra[13] combines the principal inertia into two dependent variables α and β as follows:

$$\alpha \equiv (J_y - J_z)/J_x = -0.282 \tag{11}$$

$$\beta \equiv (J_y - J_x)/J_z = -0.135 \tag{12}$$

The stability properties of the gravity gradient torque may then be conveniently represented by the plot in Fig. 7. Without going into detail, mechanical systems with inertia properties in the lightly shaded "Lagrange Region" are neutrally stable; i.e., the minor axis of inertia will align with the geodetic nadir vector and precess with it. The dark "DeBra-Delp Region" is stable only if the system has

Table 1 GADACS inertia properties

$J_x = 305.3$ kg m^2
$J_y = 258.7$ kg m^2
$J = 344.7$ kg m^2

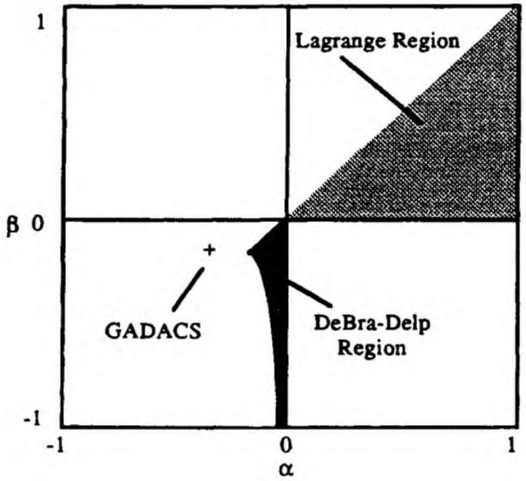


Fig. 7 Gravity gradient stability regions.

no energy damping. Because the SPARTAN mass properties lie outside these regions, it is concluded that gravity gradient torques alone are destabilizing and would reorient the spacecraft with the GPS antenna array looking along the local horizontal without active control.

B. Aerodynamic Moment

The aerodynamic torque is represented in Fig. 8. It is caused by a difference between the location of the center of pressure, where the resultant drag force acts along the negative velocity ("ram") vector and the center of mass, about

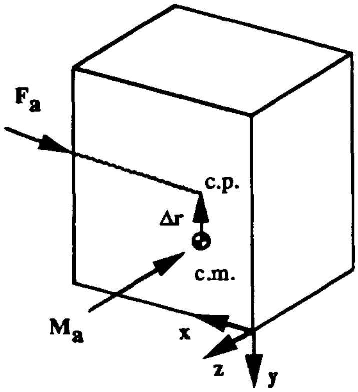


Fig. 8 Aerodynamic moment.

which moments are taken. The GADACS center of pressure (the center of area) is almost directly along the vehicle z axis:

$$\Delta \boldsymbol{r} = 18.8\, \boldsymbol{k} \text{ cm } \equiv z_b \tag{13}$$

$$\boldsymbol{M_a} = \Delta \boldsymbol{r} \times \boldsymbol{F_a} = M_a \hat{\boldsymbol{z}}_b \times (-\hat{\boldsymbol{v}}_b) \tag{14}$$

where

$$M_a \equiv \max |M_a| = 3.39e - 4 \text{ N m} \tag{15}$$

and the ($\hat{}$) notation signifies the unit vector operation.

The velocity of the satellite is easily represented in the orbit reference frame for the circular GADACS orbit as follows:

$$\boldsymbol{v_o} = |v| \boldsymbol{i_o} \tag{16}$$

Then, following the 321 Euler sequence transformation

$$\hat{\boldsymbol{v}}_b = (T_{bo})^T \hat{\boldsymbol{v}}_o = (c\psi c\theta)\boldsymbol{i} + (c\psi s\theta s\phi - s\psi c\phi)\boldsymbol{j} + (c\psi s\theta c\phi + s\psi s\phi)\boldsymbol{k} \tag{17}$$

where c and s represent cosine and sine, respectively. Thus, the aerodynamic torque in the body frame is:

$$\boldsymbol{M_a} = -M_a[(c\psi s\theta s\phi - s\psi c\phi)\boldsymbol{i} + (c\psi c\theta)\boldsymbol{j}] \tag{18}$$

$$\boldsymbol{M_a} \approx -M_a[\psi \boldsymbol{i} + \boldsymbol{j}] \tag{19}$$

The linearized vehicle equations of motion may then be determined by constraining force equilibrium

$$\dot{\boldsymbol{H}}_c - \Sigma\, \boldsymbol{M}_{\text{ext}} = 0 \tag{20}$$

which yields the following roll/yaw coupled equations:

$$\begin{aligned} J_x\ddot{\phi} + 4\, n^2(J_y - J_z)\phi + (-J_x + J_y - J_z)n\, \dot{\psi} + M_a\psi &= M_{xc} \\ J_z\ddot{\psi} + n^2(J_y - J_x)\psi + (J_x - J_y + J_z)n\dot{\phi} &= M_{zc} \end{aligned} \tag{21}$$

and in pitch:

$$J_y\ddot{\theta} + 3n^2(J_x - J_z)\theta = -M_a + M_{yc} \tag{22}$$

where M_{xc}, M_{yc}, and M_{zc} are the control torques that will be applied.

C. System Natural Response

Before progressing to the design of the control system, the dynamic behavior of the uncontrolled equations of motion should be examined. It is interesting to note that the pitch equation is decoupled from the roll/yaw system. While this result is expected for a gravity gradient disturbance, the decoupling holds in the presence of aerodynamic drag (for small motions). This property arises from the fact that the aerodynamic moment arm lies along the z-axis of the vehicle reference frame.

By completing the Laplace transform of the roll/yaw equations of motion (assuming zero initial conditions and small deflections), the following representation is obtained:

$$\begin{bmatrix} s^2 + 4\alpha n^2 & (\alpha - 1)ns + \rho \\ (1 - \beta)\, ns & s^2 + \beta n^2 \end{bmatrix} \begin{bmatrix} \phi \\ \psi \end{bmatrix} = \begin{bmatrix} M_{xc}/J_x \\ M_{zc}/J_z \end{bmatrix} \tag{23}$$

where α and β are defined above and $\rho \equiv M_a/J_x$.

The effect of aerodynamic torque on the uncontrolled system may be examined by placing the characteristic equation into Evans form vs ρ:

$$s^4 + (3\alpha + \alpha\beta + n^2)n^2 s^2 + 4\alpha\beta n^4/(1 - \beta)\, ns = \rho \tag{24}$$

When J_y is the maximum moment of inertia, the numerator may be factored as follows:

$$(s^2 + \omega_1^2)(s^2 + \omega_2^2)/(1 - \beta)ns = \rho \tag{25}$$

This relation has the root locus shown in Fig. 9. Without the aerodynamic torque ($\rho = 0$), the system is characterized by the unstable gravity gradient poles. For any $\rho > 0$, the system is unstable. The roots given by the worst case aerodynamic torque of M_a are represented by the asterisks. The dominant unstable pole for this system has a time constant of 770 seconds. Therefore, it takes about 13 minutes for a small error to grow to 2.7 times its initial value. For more typical values of aerodynamic torque ($<M_a$), the time constant is longer.

Because the growth of the instability is slow compared to the designed controller bandwidth (~0.01 rad/s, or approximately 0.1 times orbit rate ≈ 9 min), the instability may be represented as a disturbance torque, and each axis of the controller may be treated as a simple double integrator plant. It is desirable to keep the plant model as simple as possible because of the nonlinear elements of the control system described in the following section.

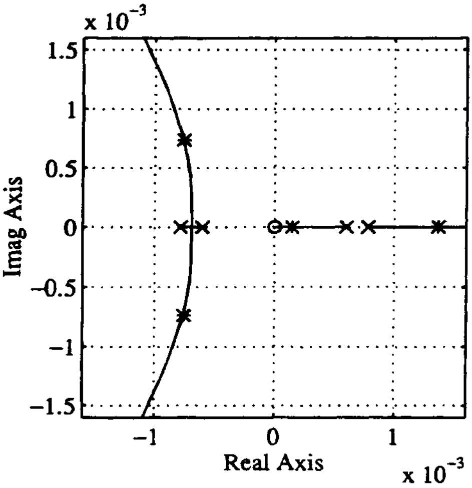


Fig. 9 Root locus for aerodynamic moment.

V. Control Design

The GADACS mission will be one of the first flight tests to quantify the accuracy of the GPS attitude solutions from experimental measurements. The expected in-flight accuracy of this device for a 1-m antenna separation is believed to be less than 0.5 deg at a 1-Hz sample rate. During the last third of the mission, the GPS solutions will be used as sensor inputs to the attitude control system; because this part of the mission is a demonstration of GPS-based attitude control, the control system will approximate several typical spacecraft pointing applications.

Because the GPS attitude solutions are relatively noisy, and the noise characteristics are not well known, and because of a very limited fuel budget, the controller requirements are to achieve reasonable pointing performance without excessive actuation in the presence of noise. Furthermore, in the event of loss of GPS attitude or significant discrepancies between the GPS attitude solutions and the inertially derived measurements, provisions have been made to use the gyro-sensed attitude for control.

A. Control Loop Description

A conceptual block diagram of the GADACS control system is shown in Fig. 10. The continuous plant is sampled by the GPS attitude receiver at an update rate of 1 Hz. Failure detection and correction logic (described in Ref. 3), chooses between the GPS solution and the inertially derived measurement of the vehicle attitude as an input to the plant estimator. This prediction estimator is designed to have settling qualities of the same natural frequency as the controller (~0.01 rad/s), and estimates both the vehicle attitude and Euler rates in roll, pitch, and yaw. These quantities are expressed in quaternion format and compared to the commanded attitude to produce an error quaternion, which is then converted into single-axis error signals that are provided to the controller. The controller is a single-axis position and velocity state feedback controller with a bandwidth of 0.01 rad/s. The command signal (with nonlinearities) is also provided as an estimator input (not shown).

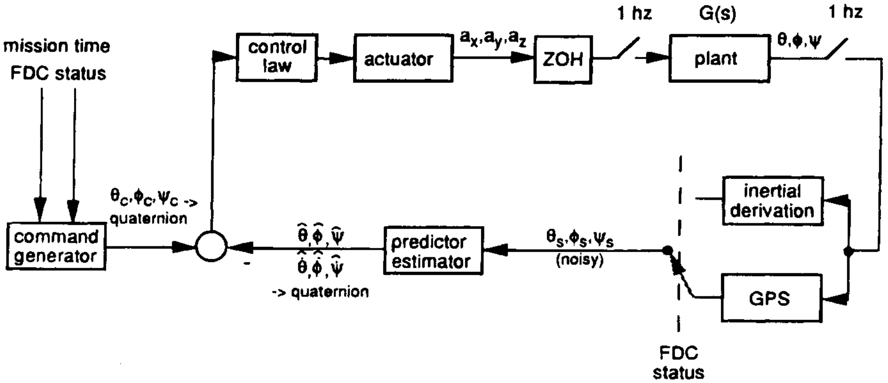


Fig. 10 Conceptual block diagram of GADACS control loop.

The control system has four nonlinearities that must be considered in the design. The control signal drives a nonlinear gas jet actuator, which is a pulse frequency-modulated system with dead zone, saturation, and rate limit. The pulse frequency modulation provides 25 ms impulses of gas calibrated to provided 0.17 deg/s^2 acceleration in a single body axis. The dead zone is designed to provide the proper fuel consumption limit cycle in the presence of sensor noise. Saturation occurs when the pulse frequency modulation reaches a full on state, which provides a constant (100% duty cycle) acceleration of 0.17 deg/s^2. A rate limit of 0.5 deg/s is implemented to prevent saturation of the DRIRU-II gyro.

For design and simulation purposes, this nonlinear system may be modeled as a linear gain with dead zone, saturation, and rate limit. Figure 11 demonstrates that this approximation is reasonably accurate when compared to the actual full nonlinear system. The effective acceleration of pulse-modulated system distributed over 1 s is compared to the simplified nonlinear model.

B. Simulation Results

Single-axis simulations were developed to demonstrate the performance of the control system. The GPS attitude solution noise was modeled as a Gaussian random process with statistically uncorrelated samples of 0.3 deg rms standard deviation. This noise level is approximately what was experienced on the RADCAL mission[2]; because of longer antenna baselines and a more favorable multipath environment, the actual noise experienced by GADACS should be less, but this level was used in the design to be conservative. In actuality, the noise is spectrally colored by multipath, but this effect was not modeled.

Environmental torques were modeled as worst-case steady-state torques acting in the same orientation. Because the controller bandwidth is high relative to the destabilizing time constant of the disturbances, no attempt was made to estimate the disturbances.

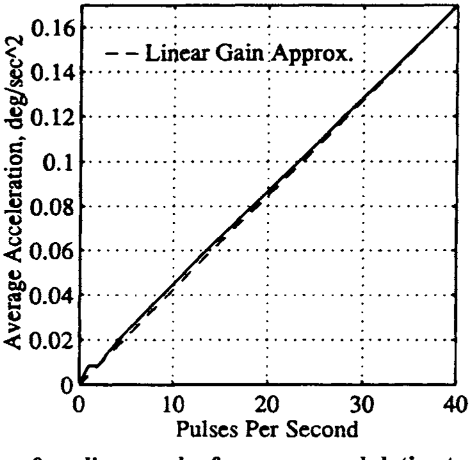


Fig. 11 Comparison of nonlinear pulse frequency modulation to linear gain approximation.

The main determinants of performance that were used in evaluating the design were controller bandwidth, steady-state error, dead zone size, and limit cycle frequency. Of these, the latter proved to be the main requirement on the design: the limit cycle frequency had to be very low, with a period greater than one-tenth orbit rate, for the design to remain within its limited actuation budget. Furthermore, the limit cycle frequency needed to be conservatively selected, given the uncertain noise characteristics of the sensor.

During the GADACS control portion of the mission, the controller will employ two types of motion, inertial hold and rate control. There will also be a limited number of inertial step and settle commands. The dead zone size was designed to achieve the proper limit cycle frequency and to reject spurious actuation caused by sensor noise. Partially because of the conservative noise characteristics used and mainly as a result of the limited fuel budget, a dead zone size for this system was selected as plus or minus 2 deg. Figure 12 demonstrates that over the course of a simulated orbit, no false actuation was produced with this deadband size for the input sensor noise. This result may be compared with a smaller deadzone of 0.5 deg in Fig. 13. No improvement in accuracy is achieved while fuel is wasted as the controller responds to the sensor noise. The dead zone is needed to keep the high-bandwidth actuator from responding to the noise in the lower-bandwidth sensor.

Figure 14 shows the design with no input, but with nonzero initial conditions to excite the limit cycle. The pitch plot shows the command input, and the sampled and estimated states. The limit cycle frequency of about six revolutions per orbit is clearly evident. The control effort required to maintain the limit cycle is reasonably small. Studies were performed to demonstrate that the limit cycle remained within the overall gas budget for up to twice the expected sensor noise.

The pitch axis, nadir-pointing profile is shown in Fig. 15. This is a rate-controlled mode with the pitch rate input equal to orbit rate (360 deg/5500 s).

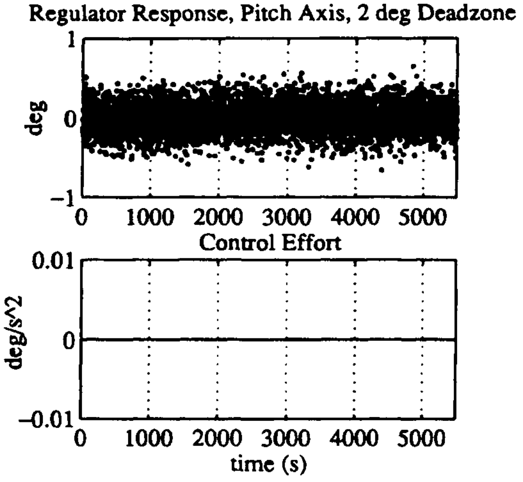


Fig. 12 2-deg dead zone reduces false actuation caused by sensor noise (estimator included).

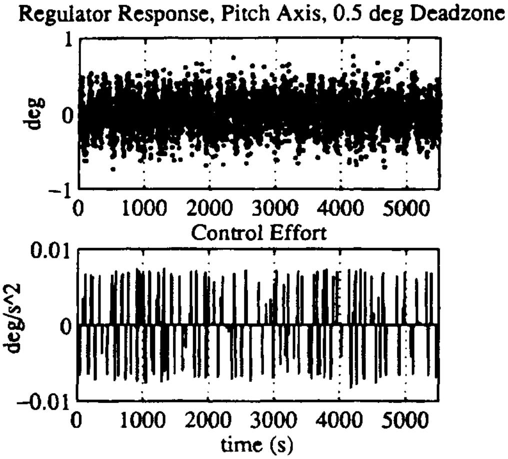


Fig. 13 0.5-deg dead zone responds to sensor noise without improving performance (estimator included).

The design is seen to accelerate to orbit rate and remain there without excess actuation once the response has settled.

The system response to a large (45 deg) step command is shown in Fig. 16. The 0.5-deg/s rate limit is seen during the first 80 s of the step. The saturation of the control system at the rate limit causes a slightly slower and more lightly damped response than in the unsaturated case, nonetheless, the performance is satisfactory.

VI. Conclusion

The design of an attitude-control system for the GADACS mission has been presented as an example of how GPS technology can be applied to spacecraft

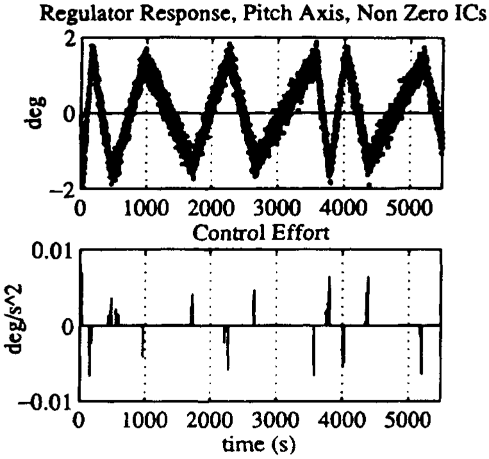


Fig. 14 Nonzero initial conditions excite limit cycle.

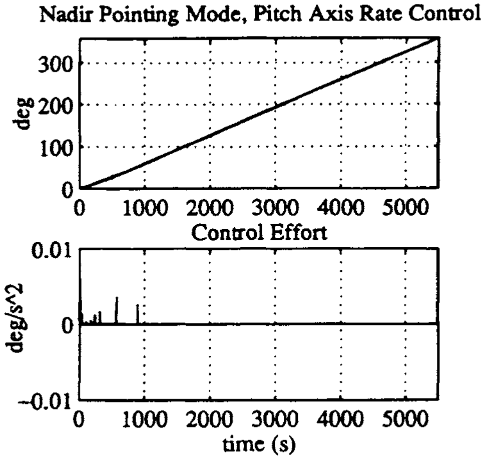


Fig. 15 Nadir-pointing mode using pitch orbit rate control.

closed-loop control. Its performance has been shown to be satisfactory with respect to fuel consumption and limit cycle frequency in the presence of sensor noise. The attitude determination accuracy of this system is expected to be less than 0.5 deg rms, and the controller accuracy is expected to be about 2 deg. Provisions have been made to allow sensor inputs to the controller from the precision inertial system in the event that GPS measurements are unavailable or unacceptable for spacecraft attitude control.

Every attempt has been made to produce a reliable control system that has good performance; however, it is important also to provide a conservative design, given the lack of flight experience for the GPS sensor. Until more space flight data are obtained, the best design is one that is fairly robust in the face of system

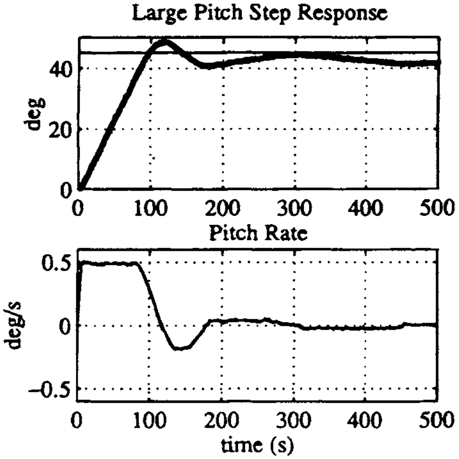


Fig. 16 Controller response to large step input.

uncertainty. As the heritage of GPS attitude determination grows, more aggressive designs can be implemented that provide the best possible performance for the system.

Cost effective subdegree attitude control is now possible in space using GPS carrier–phase-based attitude determination. This technology may soon provide an acceptable attitude sensor for many types of LEO missions. As applications grow in number, fabrication costs for space-qualified GPS receivers will be further reduced and manpower costs associated with custom designed spacecraft will be decreased. Attitude determination and control in space using GPS is an exciting new technology development that should experience substantial activity and growth over the next several years.

Acknowledgments

The author gratefully acknowledges the assistance of Penina Axelrad of the University of Colorado. The many helpful contributions of Clark Cohen and Bradford Parkinson of Stanford University; Trimble Navigation, Ltd.; and NASA Goddard Space Flight Center, including Frank Bauer, Jon McCullough, Jim O'Donnell, and the rest of the GADACS and SPARTAN design team have greatly aided in this research.

References

[1]Cohen, C. E., and Parkinson, B. W., "Expanding the Performance Envelope of GPS-Based Attitude Determination," *Proceedings of ION GPS-91* (Albuquerque, NM), Institute of Navigation, Washington, DC, Sept. 16–18, 1991, pp. 1001–1012.

[2]Cohen, C. E., Lightsey, E. G., Parkinson, B. W., and Feess, W. A., "Space Flight Tests of Attitude Determination Using GPS: Preliminary Results," *Proceedings of ION GPS-93* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 22–24, 1993, pp. 625–632.

[3]Axelrad, P., Chesley, P. B., Comp, C. J., and Ward, L. M., "GPS Based Attitude Determination," FY 93–94 TR II to the Naval Research Laboratory, Oct. 1994.

[4]Bauer, F. H., Lightsey, E. G., McCullough, J., O'Donnell, J., and Schnurr, R., "GADACS: A GPS Attitude Determination and Control Experiment on a SPARTAN Spacecraft," IFAC Conference, Palo Alto, CA, Sept. 1994.

[5]Bauer, F. H., Lightsey, E. G., McCullough, J., O'Donnell, J., Schnurr, R., Class, B. F., Jackson, L., and Leiter, S., "Pre-Flight Testing of the SPARTAN GADACS Experiment," *Proceedings of ION GPS-94* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 1994, pp. 1233–1241.

[6]Uematsu, H., and Parkinson, B. W., "Antenna Baseline and Line Bias Estimation Using Pseudolites for GPS-Based Attitude Determination," *Proceedings of ION GPS-94* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 1994, pp. 717–726.

[7]Axelrad, P., Ward, L. M., "On-Orbit GPS Based Attitude and Antenna Baseline Estimation," ION Technical Meeting, San Diego, CA, Jan. 1994, pp. 441–450.

[8]Lightsey, E. G., Cohen, C. E., Feess, W. A., and Parkinson, B. W., "Analysis of Spacecraft Attitude Measurements Using Onboard GPS," 17th Annual AAS Guidance and Control Conference, Keystone, CO, Feb. 1994, pp. 521–532.

[9] Lightsey, E. G., Cohen, C. E., and Parkinson, B. W., "Application of GPS Attitude Determination to Gravity Gradient Stabilized Spacecraft," AIAA GNC, AIAA Paper 93-3788, Monterey, CA, Aug. 1993.

[10] Kruczynski, L. R., Li, P. C., Evans, A. G., and Hermann, B. R., "Using GPS to Determine Vehicle Attitude," *Proceedings of ION GPS-89* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 1989, pp. 163–171.

[11] Schuler, B., "Impulse Budget: SPARTAN 204," TM to the SPARTAN Project, NASA/GSFC, Code 740, Oct. 1994.

[12] Hughes, P. C., *Spacecraft Attitude Dynamics,* Wiley, New York, 1986, pp. 281–346.

[13] DeBra, D. B., and Delp, R. H., "Rigid Body Attitude Stability and Natural Frequencies in a Circular Orbit," *Journal of Astronautical Sciences,* 1961.

Chapter 17

GPS for Precise Time and Time Interval Measurement

William J. Klepczynski*
United States Naval Observatory, Washington, DC 20392

Introduction

THE Global Positioning System (GPS) has quickly evolved into the primary system for the distribution of Precise Time and Time Interval (PTTI). This is true not only within the Department of Defense (DOD) but also within the civilian community, both national and international. The users of PTTI are those who maintain and distribute time (epoch) to better than one-millisecond (1 ms) precision and/or accuracy and time interval (frequency) to better than one part in ten to the ninth (1×10^{-9}). The GPS is very effective not only in meeting these modest requirements of the PTTI community but also meeting more stringent ones, such as synchronizing clocks to tens of nanoseconds over large distances.

It is not surprising that this is the case. As with all navigation systems, the heart of the GPS is a clock. In the GPS, it controls the transmission of the navigation signals from each satellite and is an integral part of the ground monitor stations. This relationship between clocks and navigation is not unique. It goes back to the eighteenth century when John Harrison (1693–1776) developed his famous clock.[1] Harrison's clock solved the longitude problem for the Royal Navy by allowing a ship to carry Greenwich time with it to sea. The navigator then determined his own local time. The difference between the navigator's local time and the Greenwich time, which he was carrying with him, was his longitude difference from Greenwich. The GPS NAVSTAR satellites are similar to the Royal Navy H.M.S. Deptford. They carry a standard reference time onboard. The navigator then uses the difference between his local time and the reference time onboard the satellite to help him determine his position.

The importance of the GPS to the PTTI community can be neither understated nor underestimated. The GPS is and will be the primary means by which time, that is Universal Coordinated Time, U.S. Naval Observatory [UTC(USNO)], the

This paper is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

*Department Head, Time Service Department.

time scale maintained at the U.S. Naval Observatory and the reference for all timed DOD systems, will be distributed within the DOD. The GPS provides time in the one-way mode (OWM) easily to a precision and accuracy of 100 ns in real-time. With a modest amount of care, it is possible to reach 25 ns. In the OWM, the GPS is considered to be akin to a clock on the wall. The output from the receiver provides time as if looking at a clock on the wall. In addition, the OWM also allows the user to determine the difference between a local clock and UTC(USNO) or GPS time. Corrections can be applied to the local clock, in real time or after the fact, so that it can be set on time to UTC(USNO) within the specifications of the system.

Through the GPS, PTTI users can also compare clocks in the common-view mode (CVM) over large distances to a precision and accuracy better than 10 nanoseconds. In the CVM, two users make measurements of their local clock with respect to the same GPS satellite at the same instant of time. If a user differences the values obtained at each site, he or she can determine the offset between the clocks at each site. However, this method requires the exchange of data by at least one of the participants.

The melting-pot method (MPM), which is similar to the OWM and requires an exchange of data as with the CVM, also allows clocks at remote sites to be synchronized and, more importantly, to be steered. In the MPM, a control station determines both the remote clock offset and rate from GPS time or UTC(USNO) and its own clock offset and rate from GPS time or UTC(USNO) by some form of regression to the observations of as many satellites as possible during the day. By comparing the two clock offsets and rates with respect to GPS time or UTC(USNO), corrections to the remote clock can be estimated. Then, corrections to the remotely located clock can be sent via a dial-up modem at any desired time. This last mode has the advantage of allowing automatic operation,[2] and it is not dependent upon any one satellite.

The ability to use the GPS in different modes to derive timing information ensures its prominence as a critical contributor to all timed systems. However, a word of caution is necessary. Prudent systems engineering requires that adequate and alternate back-up systems for PTTI be factored into the overall design of the system. *This point must be emphasized.*

I. Universal Coordinated Time

Universal Time (UT) represents a family of time scales based on the Earth's rotation on its axis. It is an important reference for navigation.

Early forms of time keeping employed time as indicated by the sundial or the apparent solar time. Ptolemy (c. 100–178 A.D.) noted the irregularity of the solar day and defined mean solar time by assuming a mean movement of the sun relative to an observer on Earth. In this way, a clock and the mean solar time were in approximate agreement. The difference between apparent solar time and mean solar time varies with the seasons and is approximated by the "Equation of Time," which in centuries past, was sometimes printed on sundials. Because the Earth is tilted on its axis by 23°.45 with respect to the ecliptic (plane of revolution of the Earth about the sun), the apparent rotation rate of the sun about the Earth is not constant throughout the year. Furthermore, the Earth's orbit around the sun is not perfectly circular. The time offset from the "Equation of Time" varies by roughly 16 min, and the two times are approximately equal four times a year—at the middle of April and June and at the end of August and December. The slight ellipticity of the Earth's orbit also causes a variation of the apparent solar day by approximately 4.7 s. We can make the appropriate correction for these effects to obtain mean solar time, and if this correction is made at the Greenwich meridian in England, we have UT0 the first UT scale. More precisely, UT0 is based on a mathematical expression for the right ascension of the fictitious mean sun and the clock time-of-transit of any celestial object with known position by an observatory yields UT0 after correction for longitude, aberration, parallax, nutation, and precession.[3] However, UT0 is not strictly uniform.

The coordinates of an observatory used to generate UT0 are subject to small changes caused by slow movements of the Earth rotation axis, called polar variation. Universal Time One (UT1) is a true navigator's time scale, which has been corrected for polar variation (PV). Universal Time One is not uniform because of small changes both seasonal (SV) and irregular and some unpredictable changes in the Earth's rotation rate. A further smoothing is performed to remove the SV to form UT2. Universal Time Two (UT2) still has small, unpredictable (10^{-8}) variation and a long-term drift (10^{-10}/year) effect. Universal Times One and Two are stable to within approximately 3 ms in a day.

With the availability of a large number and different kinds of high-precision atomic clocks, atomic time scales now exist with time stabilities better than 1 part in 10^{-14} from year to year. The accuracy of time measurement now exceeds that of any other physical measurement. Universal Coordinated Time is an atomic clock time scale coordinated by the Bureau International de Poids et Mésures (BIPM). Prior to 1982, it was coordinated by the Bureau International de l'Heure (BIH). Universal Coordinated Time differs from a pure atomic clock time, in that it occasionally introduces leap seconds because its epoch is set to astronomical time, while its rate is set to atomic time. These leap seconds are introduced to keep an atomic time scale in approximate step with the Earth's rotation. The leap second adjustment can cause a particular minute to have 59 or 61 s instead of 60. Universal Coordinated Time is, by international agreement, kept to within 0.9 s of the navigator's time scale, UT1. Leap seconds are *usually* added or deleted on June 30 or December 31. Leap seconds have been implemented since 1972.

II. Role of Time in the GPS

The practicality of using atomic clocks in space for navigation was proven with the Navigation Technology Satellites.[4] A very stable frequency source in the satellite can ensure the stability of transmissions over several revolutions of the spacecraft. This ensures adequate tracking to update the satellite's orbit and affords sufficient predictability in the clock's performance.

The GPS Block I developmental satellites contained three rubidium clocks and one cesium clock. These atomic clocks were launched into orbit to help evaluate their long-term performance in space and their effective contribution to overall operations. Rubidium clocks normally exhibit better short-term stability than cesium clocks. They also are much cheaper. However, they are subject to larger frequency variations caused by changes in environmental conditions, and they exhibit a large drift in frequency. On the other hand, cesium clocks have better long-term stability, which tends to have a greater favorable impact on operations. Experience gained with the early Block I launches helped formulate the planning for the operational Block II satellites, which were configured to have two cesium clocks and two rubidium clocks. After reaching orbit, the Master Control Station (MCS) designates the primary clock. Usually, a cesium clock has been so designated; however, in order to get operational experience with all the types of clock in orbit, a rubidium is sometimes chosen.

The orbit determination process also determines the phase offset of the satellite clock with respect to a clock or a system of clocks that has been designated as the GPS master clock. This difference is transmitted in the navigation message as the coefficients of a quadratic expression.

The GPS control segment (CS) maintains a pair of cesium beam atomic clocks at the five monitor station (MS) sites. These clocks constitute a reasonably accurate reference for the GPS orbit determination. Presently, the entire ensemble of clocks, both those in space and on the ground, form the basis of the GPS time. Thus, the GPS time is an atomic clock time similar, but not the same as, UTC time. One marked difference is that the GPS time does not introduce any leap seconds. To do so would throw the GPS P(Y)-code receivers using the system out of lock. Thus, introducing leap seconds is out of the question for GPS. Please note that the UTC leap seconds will cause the GPS time and UTC to differ by an integral and known number of cumulative leap seconds as they are introduced. Other than the leap second effect, however, the GPS CS attempts to keep the GPS time to within 1 μs of UTC time (modulo 1s).

Historically, GPS time was kept at one of the MS. The MCS had the ability to designate any MS clock as the reference for GPS time. The orbit determination process then kept track of all satellite clocks with respect to that GPS master clock. Occasionally, when something happened to that clock, then another clock was designated as the GPS master clock and the Kalman filter was re-initialized with the states of the new clock. The MS clock plays a very important role in the GPS. The determination of each of the satellite orbits is intimately tied to the MS clock. In the orbit determination process, the measured pseudoranges of each satellite are compared to and time tagged by the MS clock. Unfortunately, the orbit determination process can not separate an error in estimated range to the satellite from an error in the clock. Therefore, in order to get a good estimate of the orbit, the MS clocks have to be very stable during the estimation period.

Because environmental conditions at the MS sites were not ideal, there were frequent jumps in frequency of the MS cesium clock designated as the GPS master clock. To improve this situation, an hardware clock ensemble was installed at the MCS.[5] This was done to demonstrate the capability of a rather stable clock system at the MCS that could give a very stable reference to the GPS. This set of clocks is sometimes referred to as the "Navy clock ensemble."

At the same time that the hardware clock ensemble was being developed, a GPS composite clock was developed[6] and put into operation. This is a software clock that averages all the clocks in the system, the ground clocks and the satellite clocks. Because of the intrinsic performance of the rubidium clocks now in orbit, they have been de-weighted in their contribution to the composite clock.

III. Translation of GPS Time to Universal Coordinated Time

The GPS time is based on atomic time, and the time broadcast from the satellite is continuous (modulo 1 s) without the leap seconds of UTC. The introduction of leap seconds would throw the P-code receivers out of lock. Because the time reference for GPS and all DOD timed systems is UTC(USNO), there must be a way to relate the GPS time to UTC(USNO). This is accomplished through the use of the coefficients A_0 and A_1, also transmitted in the navigation message. These coefficients are determined through monitoring the GPS satellites at the USNO. Data used for this purpose are transmitted over a secure line to the GPS MCS at Falcon Air Force Base, CO where the coefficients are determined and transmitted to each satellite. These coefficients give the difference between GPS time and UTC(USNO). The user navigation set or timing receiver can then easily compute the difference between the local clock driving the receiver and UTC(USNO). The algorithm defining the relationship between GPS time and UTC using the navigation data in Subframe 4 as quoted in ICD-GPS-200 is as follows:

1) Whenever the effective time indicated by the WN_{LSF} (week number) and the DN (day number) values are not in the past (relative to the user's present time), and the user's present time does not fall in the timespan which starts at DN + 3/4 and ends at DN + 5/4, the UTC/GPS time relationship is given by the following:

$$t_{\mathrm{UTC}} = (t_E - \Delta t_{\mathrm{UTC}}) \text{ (modulo 86,400 s)}$$

where t_{UTC} is in seconds and

$$\Delta t_{\mathrm{UTC}} = \Delta t_{LS} + A_0 + A_1 \, [t_E - t_{ot} + 604{,}800\,(\mathrm{WN}{-}\mathrm{WN}_t)]$$

seconds; t_E = GPS time as estimated by the user on the basis of correcting t_{sv} for factors described in paragraph 20.3.3.3.3 as well as for ionospheric and SA (dither) effects

$$\Delta t_{LS} = \text{delta time attributable to leap seconds}$$

$$A_0 \text{ and } A_1 = \text{constant and first-order terms of polynomial}$$

$$t_{ot} = \text{reference time for UTC data (reference 20.3.4.5)}$$

2) Whenever the user's current time falls within the timespan of DN + 3/4 to DN + 5/4, proper accommodation of the leap second event with a possible week number transition is provided by the following expression for UTC:

$$t_{UTC} = W\,[\text{Modulo } (86{,}400 + \Delta t_{LSF} - \Delta t_{LS})], \text{ seconds}$$

where

$$W = (t_E - \Delta t_{\mathrm{UTC}}) - 43{,}200) \text{ (modulo 86,400)} + 43{,}200, \text{ seconds}$$

and the definition of Δt_{UTC} (as given in paragraph 1) above) applies throughout the transition period. Note that when a leap second is added, unconventional time values of the form 23:59:60.xxx are encountered. Some user equipment may be designed to approximate UTC by decrementing the running count of time within several seconds after the event; thereby, promptly returning to a proper time indication. Whenever a leap second event is encountered, the user equipment must consistently implement carries or borrows into any year/week/day counts. The correction parameters to convert the GPS time broadcast by the satellite to UTC time are contained in the 24 most significant bits (MSBs) of words six through nine plus the eight MSBs of word ten in page 18 or subframe.

Performance of the GPS time reference vs UTC(USNO) is shown in Fig. 1. One year of data are plotted here. The abscissa is time in units of one day. The ordinate shows the difference, in nanoseconds, between GPS time and UTC(USNO) for that day. The GPS time can be converted to UTC(USNO) by using the transmitted coefficients. We can then difference UTC(USNO) as derived from GPS, called UTC′ (USNO), and UTC(USNO) as kept at the Naval Observatory. These differences are shown in Fig. 2 for the same period of time as that shown in Fig. 1. It is important to note that the daily averages of the difference between GPS time and UTC(USNO) did not exceed 30 ns throughout the year.

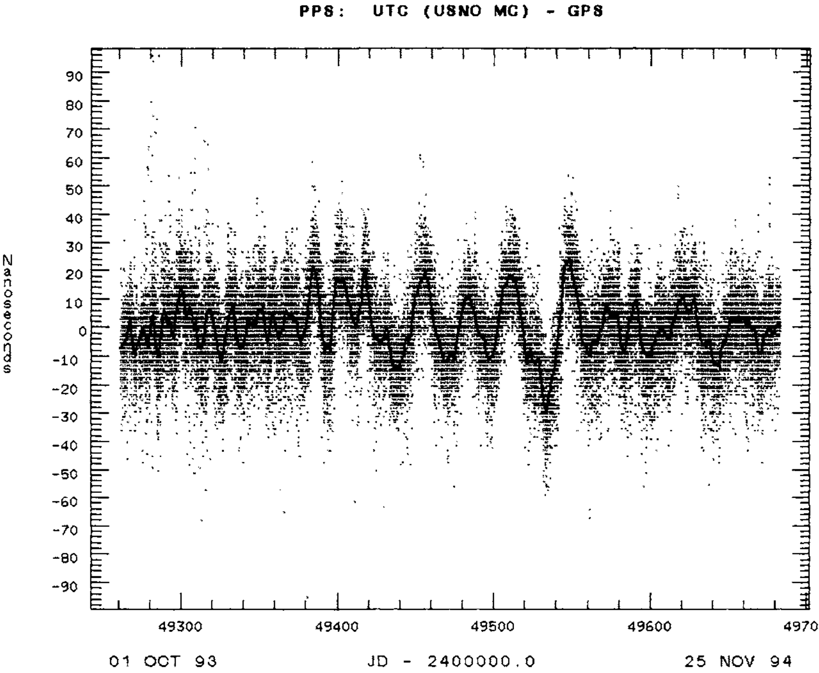


Fig. 1 UTC (USNO, MC)–GPS time over a 1-year period. One large division on the ordinate corresponds to 10 ns and on the abscissa to 100 days. The dots represent a 13-min averaged data point. A line has been drawn through the daily average of the 13-min data points.

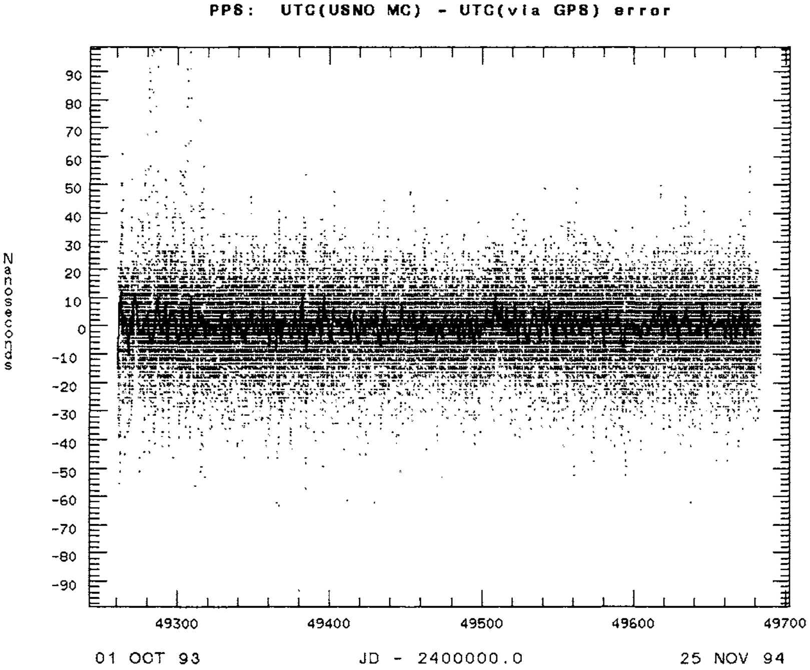


Fig. 2 UTC(USNO, MC)–UTC′ (USNO, MC via GPS) over the same 1-year period as shown in Fig. 1. UTC′ (USNO, MC via GPS) means UTC (USNO) as derived from the data contained in the navigation message. The dots represent a 13-min averaged data point. A line has been drawn through the daily average of the 13-min data points.

In addition, daily averages of the differences between UTC′ (USNO) and UTC(USNO) kept within 15 ns over the same period.

IV. GPS as a Clock in the One-Way Mode

In the computation of position, the user navigation set initially determines the difference between the local clock in the navigation receiver and each of the satellite clocks used in the fix. By application of the appropriate set of coefficients selected from the navigation message, the user can compute the difference between his local clock and GPS time or UTC (USNO). To make use of this knowledge, the receiver must then generate a timed output signal. This can be done in several ways, depending on your needs.

In most instances, a platform that has a GPS navigation unit will also be relying on that unit for time to be passed to other systems, such as, a keyed communication transceiver. Figure 3 depicts such a typical system. The timing output port of the GPS receiver is usually a serial or parallel output port, which sends the time to the system in question. The time sent is usually BCD-coded in a way that the system can understand. This, in fact, is proving to be a problem. Because of a lack of coordination, many systems have developed their own time

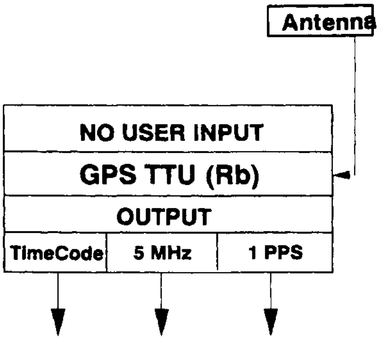


Fig. 3 Block diagram for a typical GPS time transfer unit (receiver).

Because of a lack of coordination, many systems have developed their own time code specifications. Because there is usually only one time code output port on a GPS receiver, great care must be taken in selecting which time code is ordered with the set. If time is needed for more than one system on a platform, then care must be taken in designing a distribution system that passes on the necessary codes to the other systems.

It is possible to synchronize a local clock to the output of a GPS time transfer unit (TTU) to UTC(USNO). A GPS time transfer unit is the type of timing receiver commonly used by the PTTI community. The synchronization is usually accomplished through a feedback loop of some kind and appropriate filter. A small computer keeps track of the differences between UTC(USNO) and the local clock. This information is then used by the computer to set the local clock on time. Figure 4 depicts the block diagram for such a system. In it, the user has chosen as the input to his GPS TTU a local clock driven by his local frequency standard. The rate of the local frequency standard can be adjusted, for example, by a phase microstepper, which is under computer control. The adjusted rate is then fed into a clock that can also be stepped in time by the computer. The GPS TTU sends to the computer, through a serial port, the measured difference between the local clock and UTC(USNO). The computer can then either offset the clock to bring the local time closer to UTC(USNO), or it can adjust the phase microstepper so that the adjusted rate is closer to UTC(USNO). By continually measuring and controlling this process, the local clock will eventually be set to UTC(USNO).

V. Common-View Mode of GPS

Another way in which the GPS can be used to synchronize clocks is in the CVM. This mode of measurement offers some advantages over using the GPS in the previously discussed one-way mode. In this mode, two different observers, separated by large distances, observe the same GPS satellite at the same instant of time. By taking differences of the observations made at each site, the difference between the clocks at the two sites can be obtained. However, data must be exchanged and shared between the two users. Sometimes, this can be a concern because one of the sites may not be able to transmit or share its data over a convenient link.

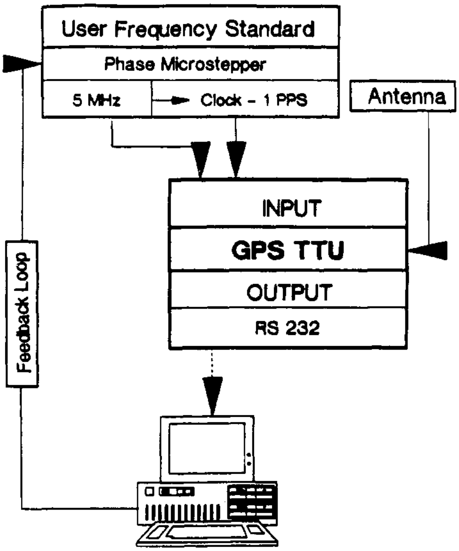


Fig. 4 Block diagram for a more sophisticated GPS time transfer unit block diagram.

The advantage of this technique is that it minimizes certain errors that might be present. Satellite clock errors are totally eliminated,[7] because they are common to both receivers. Ephemeris errors in the transmitted data are not cancelled but minimized. The amount depends upon the geometry between the two sites. Other disadvantages are that it is dependent on a few satellites, and data must be exchanged between the users. The mathematics used in this technique are very simple. We have only to subtract the values obtained at each site to obtain the differences between the two clocks at each site.

This technique is not only useful for synchronizing clocks, it could also prove to be a valuable tool in investigating ionospheric fluctuations over the two sites, as a study done in 1984 between the Tokyo Astronomical Observatory (TAO) and the U.S. Naval Observatory has shown. In this experiment, there were about 2 h of common view time between Washington and Tokyo. Observations were made as the morning terminator line crossed over between the two sites. Figure 5 exhibits the data between the two sites obtained on Feb. 15, 1984. The abscissa is time in hours. The ordinate represents the difference, in nanoseconds, between the clocks used as a reference for the local GPS time transfer units at the Tokyo Astronomical Observatory (now the National Optical Observatory of Japan) and the U.S. Naval Observatory. No discontinuities can be seen in the data between the observations that were made when both sites were in darkness and when one of the sites was in darkness and the other in daylight.

With a good amount of data, it is possible to provide for the steering of remotely located timed systems.[2] Because of the large number of satellites, adequate data can be obtained throughout the day. By smoothing over 2-day intervals, many of the fluctuations apparent in GPS data can be minimized. With the capabilities of today's rubidium and cesium clocks, this should prove more

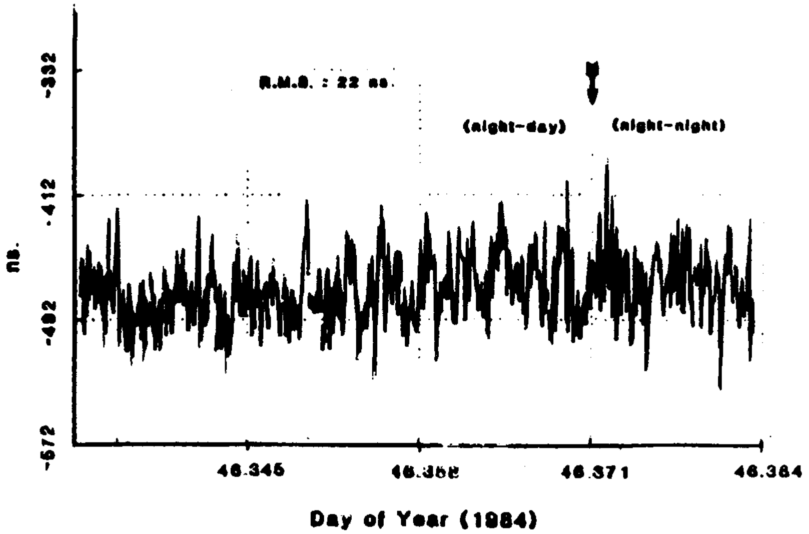


Fig. 5 Common-view GPS data, taken in 1984, between U.S. Naval Observatory (USNO) and the Tokyo Astronomical Observatory (TAO), now called the National Optical Observatory.

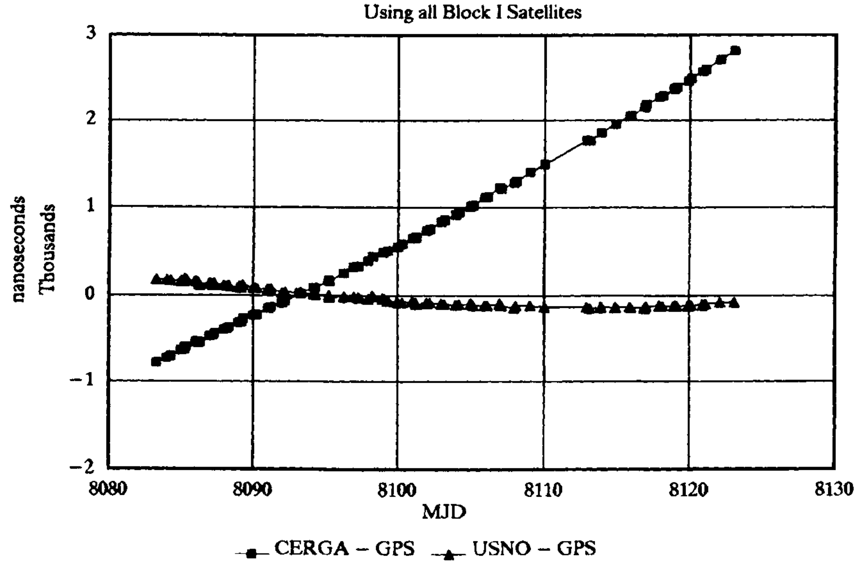


Fig. 6 Graph of data taken from Table 1, which shows the raw values of USNO–GPS and OCA/CERGA–GPS, prior to combination for GPS common-view analysis.

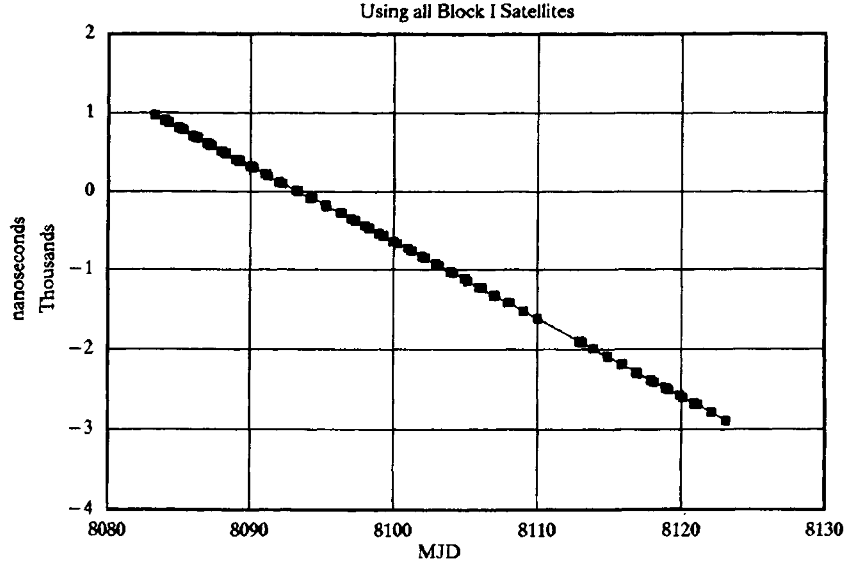


Fig. 7 Graph of the raw values of UTC (USNO)–UTC (OCA) obtained from data in Table 1 used in GPS common-view analysis. Values obtained by subtracting the data shown in Fig. 6.

than adequate to easily maintain a local time scale good to about one part in ten to the thirteenth (1×10^{-13}).

As an example of the practical application of the CVM, Table I gives a sample of GPS data obtained at two sites, the U.S. Naval Observatory in Washington, D.C. and Observatoire de la Cote d'Azure (OCA) in Grasse, France, from July 11–August 20 in 1990. Figure 6 is a plot of the differences between each station's reference clock and the GPS time. Some coordination between the two sites is required in order to ensure the greatest accuracy and precision. First, a sequence of satellites to be observed must be preselected. Factors that would come into play are geometry between the two sites, the starting time of the observations, and the length of tracking time. By selecting a satellite that passes midway between the two stations, any ephemeris errors can be minimized. However, this is not always possible to arrange. It is very important to ensure that the two sites start observing at the exact same second. In this way, all perturbations in the system are measured identically at both stations. This helps to minimize certain errors. Also, it is important to check that both stations are keeping the same time; i.e., UTC or GPS time. When one either averages or does a linear regression through the raw data, it is important that both stations have data over the exact same length of time. Some errors can bias the results because an average of data over different intervals can be different. If one differences the two sets of data, one obtains the difference between the two local reference clocks at each site. A plot of these differences is shown in Fig. 7. Figure 8 is a plot of the residuals after fitting a simple linear regression through the data exhibited in Fig. 7. The rms of the spread is about 6 ns.

Table 1 GPS Common-view data between U.S. Naval Observatory and Observatoire de l'Cote d'Azure (11 July '90–20 Aug. '90), Block I satellites

SVN[a]	MJD'[b]	USNO–GPS	OCA–GPS	USNO–OCA	Res to lin. reg.[c]
12	8083.288	−781	195	976	−2
12	8083.310	−780	189	969	−7
13	8083.985	−725	187	912	2
13	8084.007	−723	181	904	−4
12	8084.285	−705	176	881	0
12	8084.307	−699	177	876	−3
13	8084.982	−642	173	815	2
13	8085.004	−637	172	809	−2
6	8085.204	−632	165	797	5
12	8085.282	−593	199	792	8
12	8085.304	−595	193	788	6
13	8085.979	−551	164	715	−1
3	8086.090	−531	165	696	−10
6	8086.201	−541	153	694	−1
12	8086.279	−552	135	687	0
12	8086.301	−551	128	679	−6
13	8086.976	−470	146	616	−4
3	8087.088	−465	145	610	1
6	8087.199	−463	127	590	−8
12	8087.299	−438	146	584	−4
13	8087.974	−392	123	515	−8
6	8088.196	−383	115	498	−3
12	8088.274	−367	124	491	−3
12	8088.296	−370	109	479	−13
13	8088.971	−313	101	414	−12
13	8088.993	−312	101	413	−11
3	8089.082	−303	114	417	2
6	8089.193	−306	86	392	−12
12	8089.271	−269	127	396	−1
12	8089.293	−262	120	382	−13
13	8089.968	−229	96	325	−4
13	8089.990	−230	90	320	−7
3	8090.079	−230	91	321	3
6	8090.190	−215	87	302	−6
13	8090.965	−152	75	227	−5
13	8090.988	−149	82	231	1
3	8091.076	−147	78	225	3
6	8091.188	−127	77	204	−7
13	8091.985	−87	36	123	−10
6	8092.185	−57	49	106	−8
6	8093.182	25	40	15	−2
12	8093.260	22	44	22	12
12	8093.282	24	25	1	−6
6	8094.179	99	16	−83	−3
12	8094.257	90	10	−80	7
12	8094.279	88	8	−80	9
6	8095.176	182	8	−174	2

(continued on next page)

Table 1 GPS Common-view data between U.S. Naval Observatory and Observatoire de l'Cote d'Azure (11 July '90–20 Aug. '90), Block I satellites (continued)

SVN[a]	MJD'[b]	USNO–GPS	OCA–GPS	USNO–OCA	Res to lin. reg.[c]
12	8095.254	172	−7	−179	5
12	8095.276	173	−20	−193	−7
12	8096.251	266	−5	−271	10
12	8096.274	268	−12	−280	3
13	8096.971	336	−11	−347	4
12	8097.249	342	−20	−362	16
12	8097.271	346	−27	−373	7
13	8097.946	402	−34	−436	9
12	8098.246	459	−1	−460	14
12	8098.268	461	−10	−471	6
13	8098.943	498	−36	−534	8
13	8098.965	503	−32	−535	9
12	8099.243	517	−45	−562	9
12	8099.265	515	−56	−571	2
13	8099.940	558	−73	−631	8
13	8099.963	562	−76	−638	3
12	8100.263	594	−75	−669	1
13	8100.938	662	−58	−720	16
6	8101.160	680	−76	−756	1
12	8101.260	669	−94	−763	4
13	8101.935	744	−84	−828	5
13	8101.957	750	−78	−828	7
6	8102.157	761	−82	−843	11
13	8102.932	842	−95	−937	−8
13	8102.954	843	−84	−927	5
6	8103.154	865	−91	−956	−5
13	8103.929	920	−103	−1023	3
13	8103.951	925	−101	−1026	2
6	8104.151	945	−98	−1043	5
13	8104.926	1011	−97	−1108	15
13	8104.949	1018	−90	−1108	17
6	8105.149	1030	−113	−1143	2
13	8105.924	1116	−109	−1225	−5
3	8106.035	1116	−110	−1226	5
6	8106.146	1122	−110	−1232	9
13	8106.921	1227	−96	−1323	−6
13	8106.943	1229	−90	−1319	0
3	8107.032	1212	−114	−1326	1
13	8107.918	1277	−140	−1417	−4
13	8107.940	1280	−137	−1417	−1
3	8108.029	1301	−115	−1416	8
3	8109.026	1401	−117	−1518	3
3	8110.024	1490	−127	−1617	1
13	8112.904	1768	−127	−1895	2
13	8112.926	1773	−129	−1902	−2
6	8113.126	1763	−148	−1911	8
13	8113.901	1858	−135	−1993	1
13	8113.924	1863	−125	−1988	8

(continued on next page)

Table 1 GPS Common-view data between U.S. Naval Observatory and Observatoire de l'Cote d'Azure (11 July '90–20 Aug. '90), Block I satellites (continued)

SVN[a]	MJD'[b]	USNO–GPS	OCA–GPS	USNO–OCA	Res to lin. reg.[c]
13	8114.899	1958	−136	−2094	−3
13	8114.921	1958	−131	−2089	4
13	8115.896	2050	−134	−2184	4
13	8115.918	2056	−137	−2193	−3
13	8116.893	2145	−142	−2287	−2
13	8116.915	2148	−146	−2294	−7
3	8117.004	2180	−124	−2304	−9
13	8117.890	2270	−116	−2386	−5
13	8117.913	2269	−119	−2388	−4
6	8118.113	2289	−121	−2410	−7
13	8118.888	2363	−121	−2484	−6
13	8118.910	2366	−123	−2489	−9
3	8118.999	2379	−111	−2490	−1
6	8119.110	2389	−118	−2507	−7
13	8119.885	2462	−116	−2578	−3
13	8119.907	2459	−122	−2581	−4
6	8120.107	2488	−117	−2605	−8
13	8120.882	2568	−108	−2676	−4
13	8120.904	2571	−114	−2685	−11
6	8121.104	2598	−95	−2693	0
6	8122.101	2712	−81	−2793	−3
6	8123.099	2813	−81	−2894	−7

Regression Output	
Constant	785689.4
Standard error of Y Estimate	6.811978
R squared	0.999965
Number of Observations	117
Degrees of Freedom	115
X coefficient(s)	−97.0783
Standard error of coefficients	0.053525

[a] SVN: Space vehicle number.

[b] MJD': Modified Julian date.

[c] Residuals after linear regression.

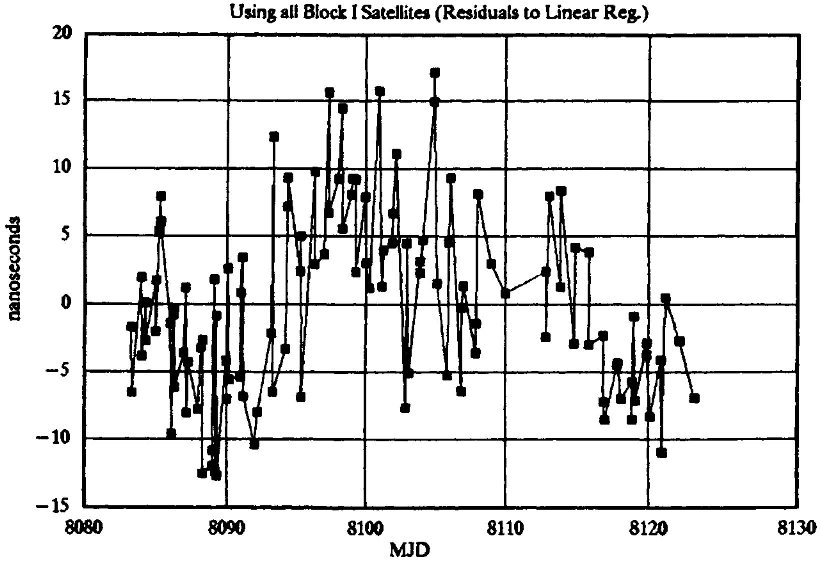


Fig. 8 Residuals with regard to a linear regression of the data shown in Fig. 7.

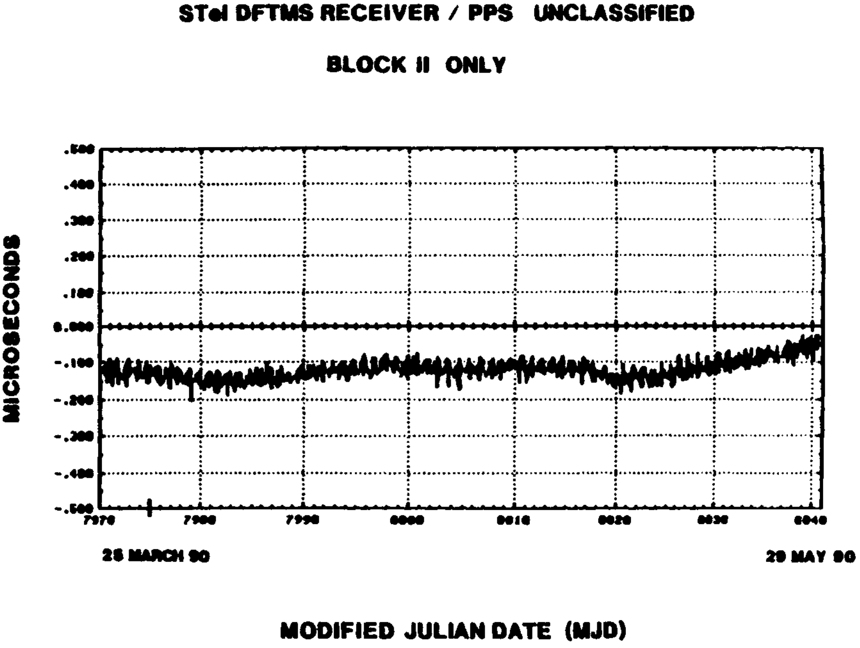


Fig. 9 UTC(USNO, MC)–GPS time over a 2-month period from which all effects of SA have been removed.

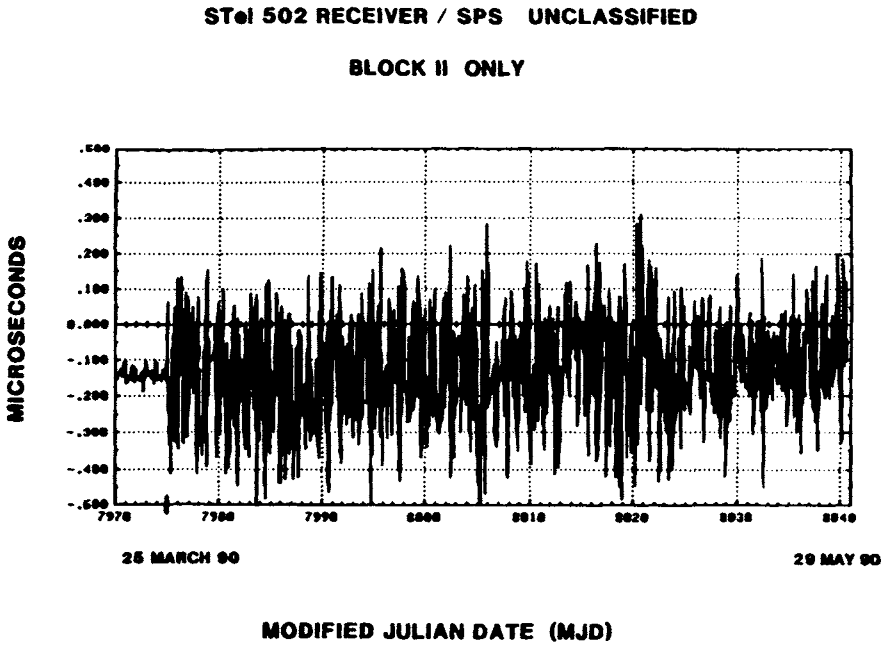


Fig. 10 UTC(USNO, MC)–GPS time over the same 2-month period as shown in Fig. 9, but the effects of SA have not been removed. It is easy to determine when SA was turned on.

VI. Melting-Pot Method

This method can also be used to synchronize clocks over widely separated distances. Unlike the CVM, which requires simultaneity of observations by both stations, the MPM only requires that each station observe as many satellites during the day that its receiver can track. This method is more robust than the CVM, because it observes significantly more satellites during the day. Therefore, it is more readily suitable to allowing the automation of steering a remotely located clock because it will not be affected by occasional gaps in data. The offset of the local clock with respect to GPS time can be ascertained by a simple regression. The offset in time and rate can than be compared with similar data obtained at another station to correct one of the clocks. In fact, it is possible to automate the process so that control of the remote clock can be done automatically when a set of prescribed limits are exceeded.

VII. Problem of Selective Availability

Selective availability (SA) will affect use of the GPS in the OWM of time transfer unless the TTU is an authorized, keyed receiver. Figure 9 shows the results of monitoring GPS with a dual-frequency TTU that has been keyed. The data are presented in a form similar to that used in earlier figures. Figure 10 shows the same data taken at the same time when SA was partially on. The

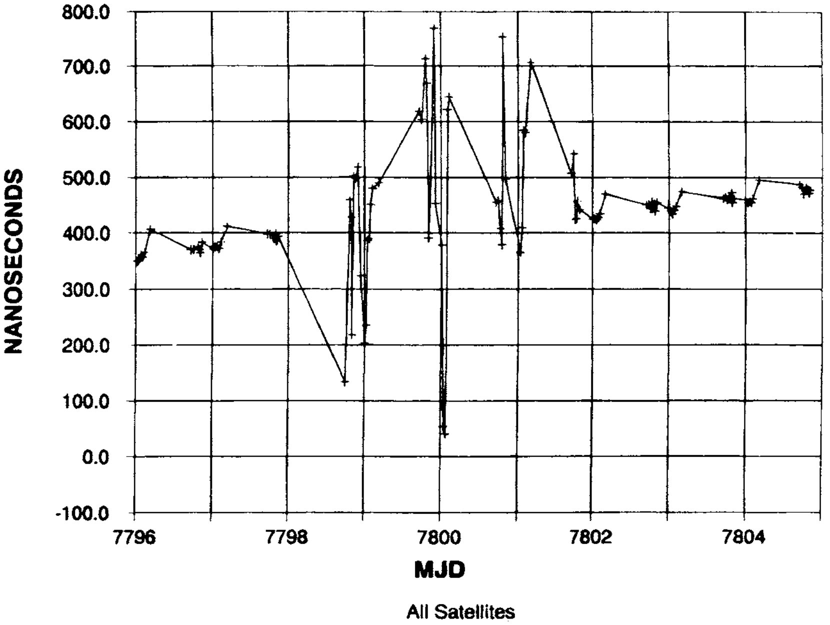


Fig. 11 Common-view comparison of two clocks during a period when SA was turned on.

degradation effects of SA on timing data is easily evident, as well as the time at which it was turned on.

The error induced in a clock that is tracking UTC derived from GPS will depend upon the level of SA. By employing some averaging techniques it is possible to minimize the affects of SA. Many of the more recent receivers on the market now track more than six satellites, thus allowing many different ways to average and smooth the data. This implies that the user clock being steered to UTC via GPS, is not updated in real time but only after a sufficient period of time has elapsed in order to smooth out the amounts of SA that are being applied.

The common-view mode of GPS can also minimize some of the effects of SA. Figure 11 shows a comparison between two different clocks via CVM GPS when SA was on. In this case, averaging will improve the results by minimizing the sudden deviations apparent in the data. The use of averaging in common-view depends upon what kind of SA is being applied. If the only form of SA being applied is clock dither, then no averaging will be necessary.[7]

IX. Future Developments

The GPS is a bright star within the PTTI community. It is now the workhorse time transfer system. It promises to be so for the next several decades. The GPS offers the user community great flexibility on how it can be used. Because of this flexibility, it has become a widely accepted and successful tool.

However, it must be cautioned that for this to continue to be so, improvements to the GPS must take place along with improvements in the PTTI field. It is obvious that a new generation of clocks will soon be developed. These clocks may have performance in the parts to the sixteenth region of stability and accuracy. There are also new techniques for time transfer over large distances that promise to exceed the current capabilities of the GPS. Therefore, for the GPS to maintain its position of pre-eminence then it, too, will have to show some progress.

There are several areas where some of the GPS subsystems can be improved, consequently improving overall system performance. The clocks at the monitor stations can be augmented with a small ensemble of cesium clocks, clock-averaging software, and independent means for comparing the clocks with those of other MSs. This would allow for the independent determination of the frequency and phase offset of the monitor station clocks from the Kalman filter process used at the MCS. In addition, newer receiver hardware can be installed at the monitor stations. There have been many improvements in receivers since the original monitor stations went into operation some 15 years ago. Because of receiver miniaturization and automation, it is now not unreasonable to think of augmenting the original set of five monitor stations to improve orbit determination.

References

[1]Gould, R. T., *John Harrison and His Timekeepers,* National Maritime Museum, 1958.

[2]Miranian, M., and Klepczynski, W. J., "Time Transfer via GPS at USNO," *Proceedings of ION GPS-91,* Institute of Navigation, Washington, DC, Sept. 1991.

[3]Winkler, G. M. R., "Timekeeping and Its Applications," *Advances in Electronics and Electronphysics,* Vol. 44, Academic Press, New York, 1972.

[4]White, J., Danzy, F., Falncy, S., Frank, A., and Marshall, J., "NTS-2 Cesium Beam Frequency Standard for GPS" *Proceedings 8th Annual Precise Time and Time Interval Applications and Planning Meeting,* Paper X-814-77-149, 1976, p. 637.

[5]Stein, S., "Improvement of CSOC Clock Ensemble Algorithm for Use at USNO," *Ball Aerospace Systems Group,* B7170-89-001, 1989.

[6]Brown, K., "The Theory of the GPS Composite Clock," *Proceedings of ION GPS-91,* Institute of Navigation, Washington, DC, Sept. 1991.

[7]Allan, D. W., and Weiss, M. A., "Accurate Time and Frequency Transfer during Common-View of a GPS Satellite," *Proceedings 34th Annual Frequency Control Symposium,* 1980, p. 334.

Chapter 18

Surveying with the Global Positioning System

Clyde Goad*
Ohio State University, Columbus, Ohio 43210

THE geodesy/geophysics and surveying communities are fortunate to have the Global Positioning System (GPS) satellites transmitting its pseudorange code on dual-frequency bands using very stable oscillators. As was discussed in Chapter 3 of the companion volume, the pseudorange code is affected by changing the phase state of the carrier by 0.5 cycles (180° or 200 gons). That is, at prescribed times, the state of the carrier is changed by 0.5 cycles, if the binary code is to be switched from a 0 to a 1 or from a 1 to a 0. These code change epochs, called chips, occur at a rate of 10.23 MHz for the P1 and P2 codes, and at a rate of 1.023 MHz for the C/A code. Should the receiver make available the difference between the phase of the transmitted signal and the phase generated by the receiver's oscillator, then these differences can be used together with the same information from other satellites being tracked by a receiver and the same satellites being tracked by other receivers. The goal of combining the one-way measures between satellites and stations is to determine, very precisely, the geometric vector (baseline) between electrical centers of two receivers' antennas. The technique used is similar to that used by electronic distance meters (EDMs), which use the (fractional) phase of the reflected signal difference to infer distance. However, unlike EDMs, the GPS signals are based on incoherent phase measures (one-way, not reflected signals) using only one or both of the two available frequencies. EDMs usually use five or so frequencies so as to determine the cycle of the reflected fractional phase without ambiguity.

Because the GPS satellite receivers can utilize only one or at most two available frequencies, some additional effort is required to determine the cycle ambiguity between oscillators, because they are not aligned. To accomplish the task of rendering the measurements in terms of quantities that will allow them to be of use, only two mathematical models are required: a model for distance traveled by an electromagnetic wave in a vacuum, and the phase change of an oscillator running with constant frequency.

*Department of Geodetic Science and Surveying.

I. Measurement Modeling

Let us first look at the physical situation, as shown in Fig. 1. The GPS satellites transmit the carrier signal continuously. Suppose the phase of the carrier that was transmitted from a GPS satellite at time t_T arrives at the ground receiver antennas at time t_R. Assume, also, that this signal travels at the speed of light through a vacuum. Thus, the distance traveled equals the time interval multiplied by the speed:

$$r = c(t_R - t_T) \tag{1}$$

This is the true range if t_R and t_T are measured by the same clock. It is pseudorange if they are measured by different clocks. The phase front of the satellite generated at transmission time, then, arrives at the receiver later at receiver sample time. This is the same assumption used to process pseudorange measurements. The actual measurement here is the difference between the satellite phase at transmission time and the receiver phase at receipt time or the following:

$$\phi_R^S(t_R) = \phi^S(t_T) - \phi_R(t_R) \tag{2}$$

where t_T is the time the carrier signal left the satellite base; t_R is the time this same signal arrived at the receiver antenna based on the receiver's clock. Superscripts refer to satellite ID; subscripts refer to receiver ID.

The left side of Eq. (2) is given as a function of receiver time here, but either t_T or t_R could be used. The choice will be dictated by how the receiver manufacturer chooses to implement Eq. (2) in hardware. This is discussed shortly.

Regardless of whether t_R or t_T is used as the reference time on the left side of Eq. (2), both t_T and t_R also appear on the right side. Now we must substitute for one of the times to obtain a corresponding phase at the time of the other.

Here, all times are expressed in terms of t_R, which is chosen because of the way data from most GPS receivers are tagged. That is, most manufacturers choose to sample the incoming phase values from all satellites being tracked at the same (received) time. This allows the manufacturers to use inexpensive oscillators. The t_R, then, represents the current state of the receiver clock at the instant phases are sampled and compared.

The connection between t_R and t_T is known from Eq. (1):

$$t_T = t_R - r/c \tag{3}$$

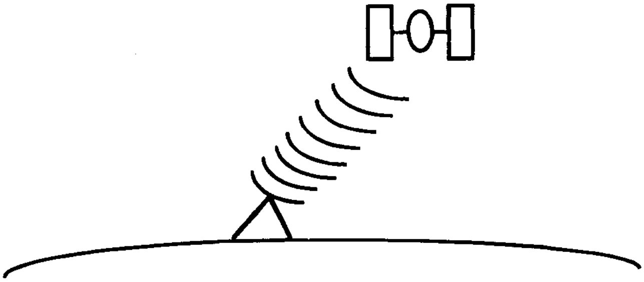


Fig. 1 GPS satellite-to-ground receiver geometry.

Substituting Eq. (3) into Eq. (2) yields the following:

$$\phi_R^S(t_R) = \phi^S(t_R - r/c) - \phi_R(t_R) \tag{4}$$

The first term on the right side of Eq. (4) can now be expanded using an ideal oscillator relation:

$$\phi(t + \Delta t) = \phi(t) + f \cdot \Delta t \tag{5}$$

Equation (5) is the same model for determining time intervals in quartz watches used by most persons today. The symbol f stands for the phase rate or frequency of the oscillator (in the satellite). We notice that the $-\ r/c$ in Eq. (4) is the Δt in Eq. (5). Thus, after substituting Eq. (5) into Eq. (4), we get the desired relation:

$$\phi_R^S(t_R) = \phi^S(t_R) - f/c\ r - \phi_R(t_R) + N \tag{6}$$

where $\phi^S\ (t_R)$ is the phase in the satellite oscillator at time t_R under the assumption of a constant running oscillator of frequency f; N is an (unknown) integer reflecting the fact that $\phi_R^s(t_R)$ measures only the fractional phase difference at time t_R or that the phase difference counter has an arbitrary integer value (i.e., only the fractional part is actually measured on the first measurement after signal acquisition).

It should be emphasized that the N is required only on the first measurement after signal lock is achieved. After lock is achieved, the phase (difference) counter counts the total (integer plus fractional) phase change from sample epoch to sample epoch. This total change in phase (integrated Doppler) continues until a loss of lock occurs.

Actually, we can directly process the phase measurement as given in Eq. (6); however, here, the generation of differenced measurements is given to show explicitly the removal of those terms not of interest to those needing position information. That is, the $\phi^S(t_R)$ and the $\phi_R(t_R)$ are of no direct interest to us. Thus, one way to remove them from the data used for positioning purposes is to generate differenced data types where the phase measurements participating in the differencing process are chosen to remove these "nuisance" variables.

Assume for now that the higher frequency L_1 from one satellite, number 6, is tracked by receivers 9 and 12. The mathematical representation of these two one-way measurements from Eq. (6) is given as follows:

$$\phi_9^6(t_K) = \phi^6(t_K) - f/c\ r_9^5(t_K) - \phi_9(t_K) + N_9^6 \tag{7a}$$

$$\phi_{12}^6(t_K) = \phi^6(t_K) - f/c\ r_{12}^6(t_K) - \phi_{12}(t_K) + N_{12}^6 \tag{7b}$$

In Eqs. (7a) and (7b), it is assumed that the two receivers sample the phase-tracking channels at exactly the same (received) time t_K. Note that $\phi^6(t_K)$ appears in both the equations.

If the two equations are differenced (7a) $-$ (7b), we get $\phi_9^6(t_K) - \phi_{12}^6(t_K) = -f/c[r_9^6(t_K) - r_{12}^6(t_K)] - [\phi_9(t_K) - \phi_{12}(t_K)] + N_9^6 - N_{12}^6$ or more simply:

$$\phi_{9,12}^6(t_K) = -f/c[r_9^6(t_K) - r_{12}^6(t_K)] - [\phi_9(t_K) - \phi_{12}(t_K)] + N_{9,12}^6 \tag{8}$$

Equation (8) gives the mathematical representation of the (between station) single difference. This type can also be used to estimate station coordinates. However, in addition to receiver (or more precisely, antenna) coordinates, we must also

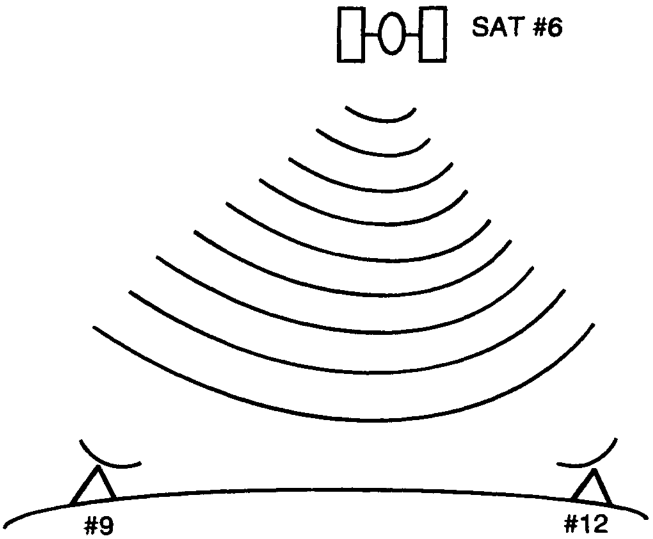


Fig. 2 Geometry involving two ground-based stations and one satellite. The figure shows the geometry of the situation. By looking at the figure, we see that the only common element between the two sampled phase measures is satellite #6.

solve for the receiver phase difference and the integer ambiguity $[\phi_9(t_k) - \phi_{12}(t_k)] + N^6_{9,12}$. This is not done here. The technique is quite similar to that used to process pseudoranges collected by a receiver during periods of no motion. If the reader is interested, the technique used to process single-difference phase measurements is given by Ref. 1.

However, generally, because of the presence of the receiver oscillator phase differences, the single differences between receiver oscillators are not of primary interest, so another differencing operation is used to remove these undesirable terms. Now we introduce another satellite (say, #18) that is also tracked by stations #9 and #12 (Fig. 3).

With the addition of satellite 18, we can generate an additional single-difference measurement. Here, the two available single differences are listed for ease of discussion.

$$\phi^6_{9,12}(t_K) = -f/c[r^6_9(t_K) - r^6_{12}(t_K)] - [\phi_9(t_K) - \phi_{12}(t_K)] + N^6_{9,12} \tag{9a}$$

$$\phi^{18}_{9,12}(t_K) = -f/c[r^{18}_9(t_K) - r^{18}_{12}(t_K)] - [\phi_9(t_K) - \phi_{12}(t_K)] + N^{18}_{9,12} \tag{9b}$$

Looking at Eq. (9a) and (9b), we see that $\phi_9(t_K) - \phi_{12}(t_K)$ is common to each. These oscillator differences between receivers can be removed through another difference operation. Thus, let us difference Eq. (9b) from Eq. (9a) to obtain the following:

$$\begin{aligned}\phi^6_{9,12}(t_K) - \phi^{18}_{9,12}(t_K) = -f/c[r^6_9(t_K) - r^6_{12}(t_K) - r^{18}_9(t_K) \\ + r^{18}_{12}(t_K)] + (N^6_{9,12} - N^{18}_{9,12})\end{aligned} \tag{10}$$

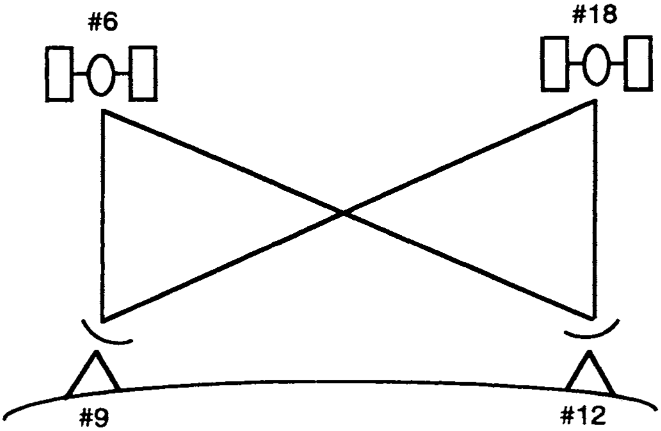


Fig. 3 Geometry involving two ground-based stations and two satellites.

Simplifying the notation as before, Eq. (10) is rewritten as follows:

$$\phi_{9,12}^{6,18}(t_K) = -f/c\left(r_9^6 - r_{12}^6 - r_9^{18} + r_{12}^{18}\right) + N_{9,12}^{6,18} \tag{11}$$

In Eq. (11), the t_K has been dropped from the r terms, because the time dependence is obvious from the t_K on the left side. The notation $\phi_{9,12}^{6,18}$ and $N_{9,12}^{6,18}$ implies that two stations (9 and 12) and two satellites (6 and 18) are involved in this "double difference" operation. $N_{9,12}^{6,18} = \left(N_9^6 - N_9^{18} - N_{12}^6 + N_{12}^{18}\right)$. Because each N value on the right is an integer, then $N_{9,12}^{6,18}$ is also an integer. The reader is now reminded of the assumptions made in the deviation of the double difference observable given in Eq. (11). They are simultaneity of reception times at receivers, perfectly constant and equal oscillator rates (frequencies) in the satellites, and that the signals from the satellites travel at the vacuum speed of light. Although none of these conditions is ever achieved exactly, small corrections can be made to the one-way measurements to achieve a high degree of compliance.

Because of the subtraction of the many common elements in Eq. (11), the double difference is not very sensitive to the (absolute) position of either receiver location, but it is sensitive to the position of one receiver relative to the other (i.e., the baseline vector). Thus, double differences are very similar to distance and angle measurements used commonly by the surveying community.

To be of use, however, we must be able to collect sufficient data to allow for the separation of geometry (baselines) and the ambiguities. For example, Fig. 4 depicts a possible history of one double difference configuration (2 satellites, 2 stations). The lower curve represents the actual geometrical contribution to the double difference phase history given by the r terms in Eq. (11). The upper curve depicts what is actually measured. Thus, the difference between the two curves is the integer ambiguity. It should be obvious that with very little data (in time) and no a priori knowledge of the baseline, there is no way to discriminate between the baseline and ambiguity. But, as time passes, there will be only one baseline that satisfies the shape or change in time of all double difference histories. Also obvious is that the greater the number of satellites

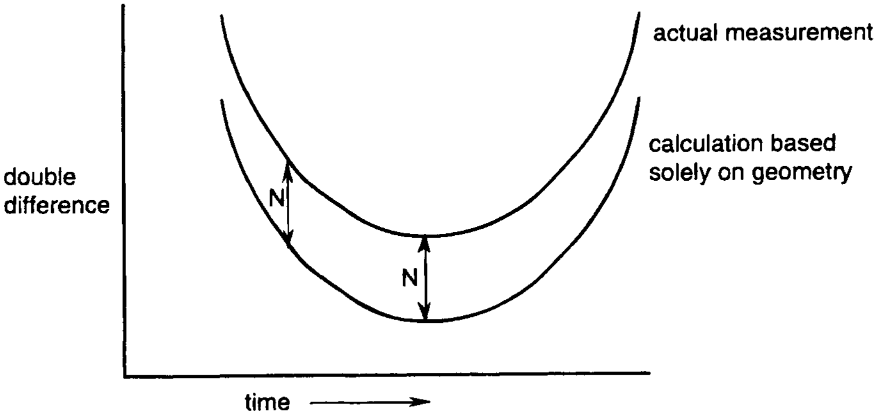


Fig. 4 A depiction of a possible double-difference measurement scenario vs what would be calculated based solely on geometry of the satellites and receivers.

tracked, the sooner the actual baseline can be identified. Once the baseline is determined unambiguously to within the order of 0.25 cycles, the *N* values can be constrained to integer values, which allows for the most desirable use of the double difference data.

Normally the baseline (vector) and ambiguities are estimated using the technique of least squares. That is, the best guess of the ambiguities and baseline are those values that minimize the sum of squares of measurement discrepancies once the estimated quantities' contributions are removed. In such implementations, we generally treat the ambiguities as real-valued parameters. These estimates, then, take on a (real) value that makes the measurement residual sum of squares a minimum. To the extent that common mode contributions to the measurements cancel, then the real-valued estimates of the ambiguities tend toward integer values. The classic case for such easy identification of integer-valued ambiguity estimates is when the baseline is short. That is, over short baselines, it is usual for those (error) sources not included in the original representation [Eq. (6)] to be removed through the differencing process [Eqs. (8) and (11)]. Such physical contributions usually canceling over short distances are errors in the refraction (tropospheric and ionospheric) and orbital errors.

Defining the concept of "short" baselines is not so easy, however. Let us consider more carefully the ionosphere, for example. The activity of the ionosphere is known to depend greatly on the 11-year cycle of sunspot activity. Therefore, when sunspot activity is low, the ionosphere is not as active, and the effect on microwave signals from GPS satellites is similar over a wider area than when the sunspot activity is greater. In 1983, when the sunspot activity was low, newly introduced single-frequency phase-measuring GPS receivers provided phase measurements that allowed for integer identification up to distances of 60 km. At the maximum of the most recent sunspot activity in 1990–1991, integers were difficult to identify, at times, over 10-km distances.

Not all possible difference combinations should be generated, however. Theoretically, only those combinations of double differences that are linearly independent offer new information to a data reduction. A linearly dependent combination is one that can be obtained by linearly combining previously used double differences. For example, consider the following possible double differences: $\phi_{9,12}^{6,18}$; $\phi_{9,12}^{6,20}$; and $\phi_{9,12}^{18,20}$. The last double difference can be obtained by a combination of the first two, as follows: $\phi_{9,12}^{18,20} = \phi_{9,12}^{6,18} - \phi_{9,12}^{6,20}$.

In other words, once $\phi_{9,12}^{6,18}$ and $\phi_{9,12}^{6,20}$ have been used, no new information is contained in $\phi_{9,12}^{18,20}$. Thus, such linearly dependent data should not be considered. If n represents the number of receivers and s the number of satellites being tracked at a data-sampling epoch, the maximum number of linearly independent combinations is $(n - 1)(s - 1)$. For the simple case of only two receivers, the generation of linearly independent data is not so difficult. However, when the number of receivers is greater than two, the task of generating the maximum number of linearly independent measurements in order to gain the maximum amount of information possible is not so trivial. Reference 2 has addressed this problem in detail.

Because there are usually several ways to combine data to form independent observables, there may be advantages of some schemes over others. Distance between receivers is one such consideration. Let us consider the case of three ground receivers (A, B, C) collecting data simultaneously as given in Fig. 5.

Here there are three possible baselines, only two of them linearly independent. Which two should be chosen? Now it is appropriate to discuss those contributions ignored in the generation of Eq. (11). These include such items as tropospheric and ionospheric refraction, multipathing, arrival time differences, orbit error, etc. Two of these unmodeled contributions are known to have errors that increase with increasing distance between receivers—orbit error and ionospheric refraction. (Tropospheric refraction does also, but only to a limit of, say, 15–50 km). Now, back to Fig. 5. Because we now realize that a more complete cancellation of unmodeled errors occurs for the shorter baseline, and thus, the use of Eq. (11) is more justified, we definitely should choose the baseline BC as one of the two independent lines. Although not so drastically different, we might as well choose AB as the other independent line, because it is slightly shorter than the line AC.

Although both orbit error and ionospheric influences are baseline length-dependent, the ionosphere causes the major degradation. Solar storms, traveling ionospheric disturbances (TIDs), day/night variations, etc. can cause large disturbances in the GPS signals. This is especially bothersome at equatorial and auroral latitudes. Moreover, these disturbances are especially prominent at the peaks of the 11-year solar cycle. Fortunately, there is one "fix" and that is to use the dispersive character of the ionospheric effect. Dispersion means frequency dependent. So, two signals transmitted at different frequencies will exhibit different

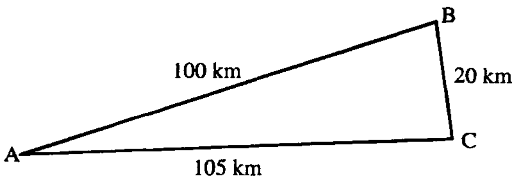


Fig. 5 Possible geographical distribution of satellite receivers.

ionospheric signatures. Actually, a better mathematical model for the GPS signals is one that recognizes a retarding of the code (group delay) and an advance of the phase—both inversely proportional to the square of the transmission frequency when expressed in distance units (say, meters).

Thus, having two measurements each at a different frequency, allows us to combine them in such a way that the ionospheric effect can be cancelled. Fortunately the GPS system was designed to transmit two different frequencies—L_1 (1575.42 MHz) and L_2 (1227.6 MHz). However, these two frequencies are transmitted only on the P codes and not on the C/A code, which is available to the civil sector. Not only is a dual-frequency receiver more expensive, it might be unable to track the dual-frequency P-code signals if antispoofing (AS) is being used.

Because the ionosphere contribution is inversely proportional to the square of frequency in range, then it is equivalently inversely proportional to frequency (to the first power) in angular units (cycles). Thus, we can now modify Eq. (6) to incorporate the ionospheric contribution as follows:

$$\phi_R^s(t_R) = \phi^s(t_R) - (f/c)r - \phi_R(t_R) + N + I(t)/f \tag{12}$$

Let us assume that we want to combine phase measurements at the L_1 and L_2 frequencies so as to remove the ionospheric terms I/f_i (where f_i stands for either the L_1 or L_2 frequency; i.e., $i = 1$ or 2); then, ϕ (no ion) $= \alpha_1\phi(L_1) + \alpha_2\phi(L_2)$ represents the "ion-free" combination. The I/f_i terms will cancel if the condition $\alpha_1/f_1 + \alpha_2/f_2 = 0$ is satisfied. In addition, another condition can be imposed, so the resulting combination is usually chosen so as to look like the original L_1 equation, but without the ionospheric term. This additional condition is given as $\alpha_1 f_1 + \alpha_2 f_2 = f_1$. These two conditions allow for a unique solution as follows:

$$\alpha_1 = f_1^2/\left(f_1^2 - f_2^2\right) = 2.5457$$

$$\alpha_2 = -f_1 f_2/\left(f_1^2 - f_2^2\right) = -1.9837$$

For the case of an "ion-free" measure, Eq. (12) should be written as follows:

$$\phi_R^S(t_R) = (\alpha_1 + \alpha_2)\phi^S(t_R) - (f_1/c)r - (\alpha_1 + \alpha_2)\phi_R(t_R) + \alpha_1 N_1 + \alpha_2 N_2 \tag{12$'$}$$

Here note that the ionosphere term I is time dependent. We have combined the L_1 and L_2 measurements to eliminate this ionospheric term—but at a price. We must now work with increased noise. Furthermore, the ambiguities are no longer integers, because the coefficients needed in the no-ion combination are not themselves integers, which destroys the integer nature of the resulting ambiguities.

Because of increased processing requirements and increased costs associated with receiver purchases, many surveyors choose to purchase or rent the less expensive single-frequency receivers (L_1 only) and try to counter the detrimental impact of the ionosphere by observing only over short baselines, as discussed earlier.

II. Dilution of Precision

Because the mathematical model of the double difference observables can be generated even before a survey is undertaken, with the anticipated data collection start/stop times and recent satellite ephemeris information (almanac), the least-squares adjustment process can be simulated that allows for a predetermination of the geometrical strength in the planned data. This has been useful in deciding the amount of time needed for a survey session. These recovered dilution of precision (DOP) values are, in essence, the very same as the position dilution of precision (PDOP) or geometric dilution of precision (GDOP) values used in navigation and described in the companion volume, but now based on the accumulated data over a survey session rather than on data collected at an instant, as is done for navigation purposes.

Two such measures can be calculated, one based on the ability to fix the ambiguities to their integer values (the "fixed" DOP values) and one that assumes that the integer ambiguities cannot be determined and must be estimated as real numbers along with baseline components (the "float" DOP value). Experience with a particular set of hardware and software techniques along with the calculated DOP values allows us to estimate the amount of data required to identify double difference integer ambiguities. Based on the techniques discussed up to this point, 0.5–1.0 h per baseline are typical.

III. Ambiguity Search

With the rapid improvement of personal (low-cost) computers, a technique introduced by Ref. 3 is now being pursued by some investigators. In essence, it is a search technique that requires baseline solutions to have integer ambiguities. Two techniques have evolved; one that searches arbitrarily many locations in a volume and one that restricts the search points to those locations associated with integer ambiguities. To put it another way, one "loops" over all locations in a volume, or one loops over possible integer ambiguity values that yield solutions within a given volume. The explanation of this technique requires only the use of Eq. (11). A sample location in space is chosen (arbitrarily). It can then be used to calculate the distances (r terms) in Eq. (11), and if it is the actual location of the antenna, then that which is left after removing the r terms should be an integer. All measurements to all satellites at all epochs will exhibit this behavior. Locations that do not satisfy this requirement cannot be legitimate baselines. The beauty of this search technique is that cycle slips (losses of lock) are not a consideration; that is, even if the ambiguity changes its integer value, such an occurrence has no impact on the measure of deviation from an integer.

The volume search technique is the easiest to envision and the most robust. A suitable search cube, say one meter on a side, is chosen and each location in a grid is tested. Initial search step sizes of 2–3 cm are reasonable. Once the best search point is found, a finer search can be performed to isolate the best fitting baseline to, say, the mm level. Although the most robust, this volume search can be time consuming. An alternative is to choose the four satellites with the best PDOP, and test only those locations found from assuming that their ambiguities are integers. That is, "loop" on a range of ambiguities rather than all locations in the test cube. Such a scheme is much faster, but can suffer if the implied test locations are in error because of unmodeled contributions to the measurements used to seed the search. Effects that can cause such errors are multipathing, ionosphere, etc.

In either case, the key to minimizing computer time is to restrict the search volume. One such way is to use differential pseudorange solutions if the pseudoranges are of sufficient quality. Here P-code receiver measurements are usually superior to those that track only the C/A codes. However, some manufacturers are now claiming to have C/A-code receivers with pseudorange precision approaching 10 cm. Of course, success can only be obtained if the initial search volume contains the location of the antenna within it. So one now must contend with competing factors. The search volume needs to be as large as possible to increase the probability that the true location can be found. However, the search volume must be small enough to obtain the estimate in a reasonable amount of time. Clearly, the better the available pseudoranges and the greater the number of satellites being tracked, the better such a search algorithm will work.

These search techniques can also be used even when the antenna is moving. However, in this case, one must assume that no loss of lock occurs for a brief time so that the search can be performed on ambiguities. This, then, allows for the different epochs to be linked through a common ambiguity value, because there is no common location between epochs of a moving antenna (unless the change in position is known, which could be the case if inertial platforms are used). As computers become even more powerful and if receivers can track pseudoranges with sufficient precision and orbits are known well enough, even baselines over rather long distances can be determined using these techniques. In the end, because of the required computer time, one probably would not use these search techniques to determine the entire path of an airplane or other moving structure, but they could be very useful in providing estimates of integer ambiguities in start-up or loss-of-lock situations.

IV. Use of Both Pseudoranges and Phase

It should now be obvious that for the most precise surveying applications, recovery of the ambiguities is required. Using the approach discussed earlier, the separation of the geometrical part (baselines) and the ambiguities requires some time to pass in order to utilize the accumulated Doppler. One major consequence of this approach is that the integer ambiguities are more difficult to identify with increasing baseline length because of the decoupling of unmodeled error sources such as tropospheric refraction and orbital errors. The same is true for the ambiguity search. With the introduction of affordable receivers collecting both dual-frequency pseudoranges and phases, this laborious approach might be laid to rest if sufficient noise reduction can occur with the tracking of the precise pseudoranges. Techniques utilizing the P-code pseudoranges are now discussed.

For some time, the ability to use readily the pseudoranges in addition to dual-frequency phase measurements provided by the ROGUE receivers designed at the California Institute of Technology/Jet Propulsion Laboratory to recover widelane phase biases has been well known.[4,5]

Here we examine the simultaneous use of all four measurements (phases and pseudoranges from both L_1 and L_2 frequencies). It is shown that the four-measurement filter/smoother can be generated numerically from the average of two three-measurement filters/smoothers. Each of the three-measurement algorithms can be used to provide estimates of the widelane ambiguities, provided that some preprocessing can be performed to reduce the magnitude of the L_1 and L_2 ambiguities to within a few cycles of zero. However, such a restriction is not required for the four-measurement algorithm.

A. Review

To aid in understanding these techniques, a review is presented using the notation of Ref. 5. First, the set of measurements available to users of receivers tracking pseudoranges and phases on both the L_1 and L_2 frequency channels at an epoch is given mathematically as follows:

$$\rho_1 = r^* + I/f_1^2 + \epsilon_{R_1} \tag{13a}$$

$$\Phi_1 = r^* - I/f_1^2 + N_1\lambda_1 + \epsilon_{\Phi_1} \tag{13b}$$

$$\rho_2 = r^* + I/f_2^2 + \epsilon_{R_2} \tag{13c}$$

$$\Phi_2 = r^* - I/f_2^2 + N_2\lambda_2 + \epsilon_{\Phi_2} \tag{13d}$$

Note, here that all the measurements in Eqs. (13a–13d) are expressed in linear units, which is different from that given earlier in Eq. (12) for phases. A simple scaling by λ_1 or λ_2 accomplishes this transformation; $\Phi = \lambda \times \phi$.

In Eqs. (13a–13d), the r^* represents the combination of all nondispersive clock-based terms; or, in other words, the ideal pseudorange; the dispersive ionospheric contribution is the I/f_i^2 term (theoretically a positive quantity) with group delays associated with pseudoranges and phase advances associated with the phases. The two phases (range) measurements include the well-known integer ambiguity contribution when combined in double-difference combinations. Finally, all measurements have noise or error terms ϵ.

Eqs. (13a–13d) can be expressed in the more desirable matrix formulation as follows:

$$\begin{bmatrix} \rho_1 \\ \Phi_1 \\ \rho_2 \\ \Phi_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & \lambda_1 & 0 \\ 1 & (f_1/f_2)^2 & 0 & 0 \\ 1 & -(f_1/f_2)^2 & 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} r^* \\ I/f_1^2 \\ N_1 \\ N_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{\rho_1} \\ \epsilon_{\Phi_1} \\ \epsilon_{\rho_2} \\ \epsilon_{\Phi_2} \end{bmatrix} \tag{14}$$

In Eq. (14), it is readily apparent that in the absence of noise, one would solve the four equations in four unknowns to recover ideal pseudorange, instantaneous ionospheric perturbations, and the ambiguities. Although the noise values on phase measurements are of the order of a millimeter or less, the pseudorange noises vary greatly from receiver to receiver. L_1 C/A-code pseudoranges have the largest noise values, possibly as high as 2–3 m. This is because of the relatively slow chip rate of 1.023 MHz. P-code chip rates are 10 times more frequent, which suggests noises possibly as low as 10–30 cm. Obviously, to determine ambiguities at the L_1 and L_2 carrier frequencies ($\lambda_1 \cong 19$ cm, $\lambda_2 \cong 24$ cm), low pseudorange noise values play a critical role in the time required to isolate either N_1 or N_2, or some linear combination of them. In a least-squares smoothing algorithm, Ref. 5 showed that the worst and best combinations of L_1 and L_2 ambiguities are the narrow-lane ($N_1 + N_2$) and wide-lane ($N_1 - N_2$)

combinations, respectively. With 20-cm pseudorange uncertainties, the wide-lane estimate uncertainty approaches 0.01 cycles; whereas, narrow-lane uncertainties are at about 0.5 cycles. These should be considered as limiting values, because certain contributions to Eqs. (13) and (14) were not included, such as multipath and higher-order ionosphere terms, with multipath being the more dominant of the two, by far.

The beauty of Eq. (14) lies in its simplicity and the ease of implementing a least-squares algorithm to obtain the wide-lane ambiguity values. Once the wide-lane ambiguity is obtained, the usual ion-free combination of Eqs. (13b) and (13d) yield biases that can be expressed as a linear combination of the unknown L_1 ambiguity and the known wide-lane ambiguity. Knowing the values of the wide-lane ambiguity makes it much easier to recover the L_1 ambiguity. However, not knowing either ambiguity, and even knowing the baseline exactly is a situation wherein it is possible the analyst will be unable to recover the integer values for N_1 and N_2.

Other factors, in addition to multipath, that could negatively influence the use of Eq. (13) would be the nonsimultaneity of sampling of pseudorange and phase measurements within the receiver or a smoothing of the pseudoranges using the phase (or Doppler) information that attempts to drive down the pseudorange noise but then destroys the relations (13a–13d). Note that theoretically no large ionosphere variations or arbitrary motions of a receiver's antenna negate the use of Eqs. (13) or (14). Thus, after sufficient averaging, wide-lane integer ambiguities can be determined for a receiver/antenna, say, involved in aircraft tracking or tracking a buoy on the surface of the sea. For many terrestrial surveys, once sufficient data have been collected to recover the wide-lane ambiguity, no more would be required, except where total elimination of the ionosphere is required, such as for orbit determination and very long baseline recoveries. For these situations, both L_1 and L_2 integers are desired, and geometry changes between satellite and ground receivers are required unless the baseline vectors are already known. The technique of using such short occupation times along with the four-measurement filter to recover wide-lane ambiguities is known as "rapid static surveying." Again, one must be aware that unmodeled multipath can be very detrimental when very short occupation times are utilized.

An example of the use of Eq. (14) in a least-squares algorithm is illustrated in Fig. 6. Here four measurements ρ_1, ρ_2, Φ_1, Φ_2 were collected every 120 s at the Penticton, Canada, tracking station. Although the integer nature of the ambiguity can only be identified after double differencing, the one-way measurements (satellite-to-station) can be smoothed separately, and the biases combined later to yield the double difference ambiguities. Figure 6 shows the difference between the linear combination involving ρ_1, ρ_2, Φ_1, Φ_2 to yield the wide-lane ambiguity on an epoch-by-epoch basis with the estimated values. The reader will notice that individual epoch values deviate little from the mean or least-squares estimate; the rms of these values is 0.06 cycles. The three-measurement combinations are discussed in the next section.

Table 1 shows the estimates of the double-difference ambiguities formed from the combination of one-way bias estimates between Canadian locations Penticton and Yellowknife, which are 1500-km apart. The integer nature of the wide-lane values is clearly seen, while the similar integer values of the L_1 and L_2 bias values

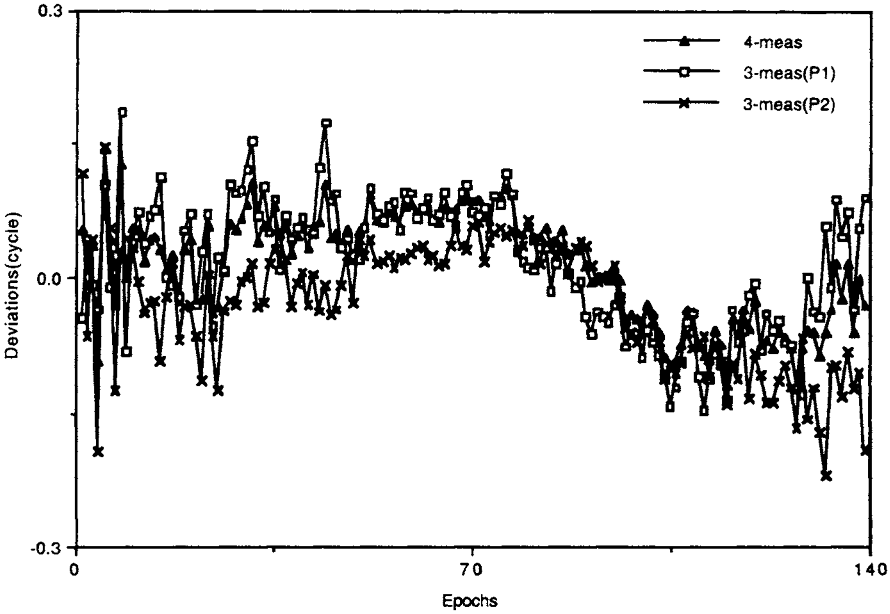


Fig. 6 Deviations from mean values of the four- and three-measurement combinations, Rogue receiver, Penticton, Canada, day 281, 1991, space vehicle (SV) 14.

Table 1 Estimated values of the N_1, N_2, and wide-lane ($N_1 - N_2$) double-difference ambiguities

Sat	Sat	N_1	N_2	$N_1 - N_2$
2	3	−0.162	−1.177	1.105
2	6	−0.284	−1.250	0.966
2	11	−0.002	−1.044	1.042
2	12	−0.539	−1.497	0.957
2	13	−0.450	−1.396	0.947
2	14	1.544	−0.542	2.086
2	15	−1.492	−2.562	1.070
2	16	0.174	−0.877	1.051
2	17	0.035	−1.015	1.051
2	18	−0.335	−1.382	1.047
2	19	0.905	−0.119	1.024
2	20	−0.214	−1.197	0.983
2	21	0.078	−0.984	1.063
2	23	−0.253	−1.316	1.063
2	24	−0.787	−2.735	1.948

cannot be identified. Clearly, in the processing steps, an integer close to the originally determined bias value has been subtracted from the corresponding phase measurements in an attempt to keep the double-difference ambiguities close to zero. This was not a requirement of the four-measurement technique, however.

B. Three-Measurement Combinations

Here the derivation of the two three-measurement combinations is presented. First, we must use the two phase measurements, Eqs. (13b) and (13d). Next, choose only one of the two pseudorange measurements ρ_1 or ρ_2. Let us choose to examine the selection of either by denoting the chosen measurement as ρ_i, where i denotes either 1 or 2 for the L_1 or L_2 pseudorange, respectively. To simplify the use of the required relations, Eqs. (13a–13d) are rewritten as follows:

$$\rho_i = r^* + I/f_1^2 + \epsilon_{\rho_i} \tag{15a}$$

$$\Phi_1 = r^* - I/f_2^2 + N_1\lambda_1 + \epsilon_{\Phi_1} \tag{15b}$$

$$\Phi_2 = r^* - I/f_2^2 + N_2\lambda_2 + \epsilon_{\Phi_2} \tag{15c}$$

The question to be answered is: What is the final combination of N_1 and N_2 after eliminating the r^* and I terms in Eqs. (15a–15c)? The desired combinations can be expressed as follows:

$$\mathrm{a}\rho_i + \mathrm{b}_{\Phi_1} + \mathrm{c}_{\Phi_2} = \mathrm{dN}_1 + \mathrm{eN}_2 + \mathrm{a}\epsilon_{\rho_1} + \mathrm{b}\epsilon_{\Phi_1} + \mathrm{c}\epsilon_{\Phi_2} \tag{16}$$

where $d = b\lambda_1$, and $e = c\lambda_2$.

To assure the absence of the r^* and I terms, the a, b, and c coefficients must satisfy the following:

$$a + b + c = 0 \tag{17a}$$

$$a/f_i^2 - b/f_1^2 - c/f_2^2 = 0 \tag{17b}$$

One free condition exists. Because it is desirable to compare the resulting linear combinations of N_1 and N_2 to the wide-lane combination, we choose arbitrarily to enforce the following condition:

$$d = d\lambda_1 = 1 \tag{17c}$$

Solving Eqs. (17a–17c) with $i = 1, 2$ yields the two desired three-measurement combinations with noise terms omitted:

$$\begin{aligned} -1.2844\rho_1 + 5.2550\Phi_1 - 3.9706\Phi_2 = {} & N_1 \\ & - 0.9697N_2, \qquad \text{for} \quad i = 1 \end{aligned} \tag{18a}$$

$$\begin{aligned} -1.0321\rho_2 + 5.2550\Phi_1 - 4.2229\Phi_2 = {} & N_1 \\ & - 1.0313N_2, \qquad \text{for} \quad i = 1 \end{aligned} \tag{18b}$$

In practice, the coefficients in Eqs. (18a) and (18b) should be evaluated to double precision. The errors in the above combinations are dominated by the pseudorange errors that depend on the receiver characteristics, as discussed earlier. However, when compared to even the most precise GPS pseudoranges, the phase uncertainties are orders of magnitude smaller. Thus, the error in the combination

(18a) in cycles is equal to 1.28 times the uncertainty of ρ_1 (in meters). Similarly, the combination (18b) is equal in cycles to 1.03 times the uncertainty in ρ_2 (in meters). As with the four-measurement combination, averaging can be used to reduce the uncertainty of the estimated combination. Also the two three-measurement combinations possess almost all the desirable characteristics as the four-measurement combination. The same restrictions also apply. For example, simultaneity of code and phase is required; multipath is assumed not to exist; and filtering of the pseudoranges that destroys the validity of Eqs. (15a–15c) is assumed not to be present.

One situation does require some consideration—the magnitudes of N_1 and N_2. That is, in the four-measurement combination, the identification of the wide-lane ambiguity is not hindered by large magnitudes of either N_1 or N_2. However, if either of the two three-measurement combinations differ from the wide-lane integers by 3% of the N_2 value, this difference could be very large if the magnitude of N_2 is large. Thus, some preprocessing is required. For static baseline recovery, this is probably possible by using the estimated biases from the individual wide-lane and ion-free phase solutions. Using these ambiguity estimates, the L_1 and L_2 phase measurements can be modified by adding or subtracting an integer to all the one-way phases so that the new biases are close to zero. With near-zero L_1 and L_2 ambiguities, the magnitude of the 0.03 N_2 deviation from the wide-lane integer should be of no consequence in identifying the integer widelane value.

Furthermore, it appears that the average of the two three-measurement combinations is equal to the four-measurement combination. This is not the case identically; however, again, with small L_1 and L_2 ambiguities, it is true numerically.

To illustrate the power in the three-measurement combinations, the data collected on the Penticton–Yellowknife baseline are used to estimate all three combinations. Table 2 shows the resulting estimates (the last column is discussed later). It is clear that all three combinations round to the same integer values. Also

Table 2 Four-measurement and two three-measurement double-difference ambiguity estimates over the Penticton–Yellowknife baseline

Sat	Sat	$N_1 - N_2$	$N_1 - 1.03N_2$	$N_1 - 0.97N_2$	$N_1 - 1.283N_2$
2	3	1.105	1.052	0.980	1.350
2	6	0.966	1.009	0.933	1.324
2	11	1.042	1.074	1.011	1.337
2	12	0.957	1.003	0.912	1.380
2	13	0.947	0.996	0.909	1.348
2	14	2.086	2.106	2.069	2.261
2	15	1.070	1.149	0.992	1.796
2	16	1.051	1.078	1.025	1.300
2	17	1.051	1.082	1.020	1.338
2	18	1.047	1.090	1.005	1.438
2	19	1.024	1.030	1.015	1.094
2	20	0.983	1.020	0.947	1.321
2	21	1.063	1.086	1.033	1.303
2	23	1.063	1.105	1.023	1.520
2	24	1.948	2.033	1.865	2.723

apparent is that the numerical average of each of the three-measurement estimates equals the four-measurement estimate. Again, this is because of the preprocessing step to ensure that ambiguities are close to zero. Figure 6 shows deviation of the one-way (satellite-station) means from the epoch-by-epoch values. The noise levels seem to be small for all the combinations. Large scatter is noted at lower elevation angles when the satellite rises (low epoch numbers) and sets (large epoch numbers). A cutoff elevation angle of 20 deg was used in the generation of Fig. 6. Moreover, an increase in deviations with the model can be seen at the lower elevation angles.[5] The obvious question is whether this is caused by multipath.

Clearly, if we have all four measurement types, the four-measurement combinations would be used. However, with very little extra effort, all three combinations can be computed, possibly helping to identify potential problems in either the ρ_1 or ρ_2 measurements.

V. Antispoofing?

Under certain assumptions about Y-code structure (AS on), a receiver can compare the two Y codes and obtain an estimate of the difference between the two precise pseudoranges ($\rho_1 - \rho_2$). For this tracking scenario, Eq. (15a) is replaced with the following:

$$\rho_{1-2} = I/f_1^2\left[1 - (f_1/f_2)^2\right] + \epsilon_{R_1} - \epsilon_{R_2} \tag{19}$$

Imposing the same restrictions as before on the coefficients a, b, and c, the following is obtained where again the error terms are ignored:

$$(\rho_{1-2}/\lambda_1 + \Phi_1/\lambda_2 - \Phi_2/\lambda_1) = N_1 - 1.2833N_2 \tag{20}$$

The recovery of $N_1 - 1.283N_2$ using differences in pseudoranges from the Penticton–Yellowknife baseline are given in the last column of Table 2. Here, assuming the magnitude of N_2 to be less than or equal to 3, the values of N_1 and N_2 seem to be identifiable in some cases. Using an orbit to recover the ion-free double-difference biases can also be of major importance for those cases where the integer values still remain unknown to within one cycle. In any event, some concern is warranted when we are required to use these measurements. Because $1/\lambda_1 = 5.25$, an amplification of the pseudorange difference uncertainty over the individual pseudorange uncertainty of $\sqrt{2} \times 5.25 = 7.42$ is present, assuming that the pseudorange difference uncertainty is only $\sqrt{2}$ larger than either the L_1 or L_2 individual pseudorange uncertainties. This is far from the expected situation, so clearly, some noisy, but unbiased, C/A-code pseudorange data are highly desirable. The usefulness of these data types when AS is operating is an open question, and no definitive conclusions can be obtained until some actual pseudorange differences and C/A-code pseudoranges are available for testing.

VI. A Look Ahead

It is clear that receivers with precise pseudorange or pseudorange differences can make the job of finding the integer ambiguities far more robust (not depending on the orbit) and easier. Assuming that measurements are always differenced for precise relative positioning, we can envision the day when precise navigation and surveying will merge using the techniques presented here. No longer must precious time be spent prior to motion of the antenna to determine the ambiguities using geometry (change). For example, for a photogrammetric mission, an airplane could begin flight long before the receivers are even powered on. Losses of lock caused by blockages by wings or the tail section create no problem theoretically to restart. Even a temporary loss of power can be accommodated in the data-reduction process. For the first time, buoys arbitrarily placed in the oceans can be tracked to the centimeter level for studies of time variation—a true open ocean tide gauge.

Surveyors with communication gear could transmit data from a master receiver (not moving) to a receiver visiting locations whose coordinates are to be determined and recover the baselines in real time while collecting data. Once the wide-lane biases are determined, the surveyor could be notified that sufficient data have been collected, coordinates could be displayed, and the surveyor could move on to the next site of the survey, turning off the power to the receiver during the motion to extend the life of the batteries. Possibly, the most beneficial use to mankind of the real-time determination of baselines would be the access to real-time monitoring of Earth motions as precursors to earthquakes, if such precursors are, in fact, present.

Many other applications can be identified, such as automatic aircraft landing, automatic steering of a dredge or piloting of a ship, very precise roadway mapping, monitoring of Earth motions for geodynamic purposes, air gravity surveys, satellite orbit determinations, railway leveling measurements, etc. Some of these notions are developed in Chapter 15, this volume.

References

[1] Goad, C. C., and Remondi, B. W., "Initial Relative Positioning Results Using the Global Positioning System," *Bulletin Géodésique,* Vol. 58, 1984, pp. 193–210.

[2] Goad, C. C., and Mueller, A., "An Automated Procedure for Generating an Optimum Set of Independent Double Difference Observables using Global Positioning System Carrier Phase Measurements," *Manuscripta Geodaetica,* Vol. 13, 1988, pp. 365–369.

[3] Counselman, C. C., and Gourevitch, S. A., "Miniature Interferometric Terminals for Earth Surveying: Ambiguity and Multipath with Global Positioning System," *IEEE Transactions on Geosciences and Remote Sensing,* Vol. 19, No. 4, 1981, pp. 244–252.

[4] Blewitt, G., "Carrier Phase Ambiguity Resolution for the Global Positioning System Applied to Geodetic Baselines up to 2000 km," *Journal of Geophysical Research,* Vol. 94 (B8), 1989, pp. 10187–10203.

[5] Euler, H. -J., and Goad, C. C., "On Optimal Filtering of GPS Dual Frequency Observations without Using Orbit Information," *Bulletin Géodésique,* Vol. 65, No. 2, 1991, pp. 130–143.

Chapter 19

Attitude Determination

Clark E. Cohen*
Stanford University, Stanford, California 94305

ALTHOUGH originally developed as a means for navigation, GPS has since been shown to be an abundant source of attitude information as well. Using the subcentimeter precision of GPS carrier phase, a receiver can determine the relative positions of multiple antennas mounted to vehicles or platforms so accurately that their orientation may be determined in real time at output rates exceeding 10 Hz. This chapter discusses the fundamentals of attitude determination using GPS. It also describes the mathematics of attitude solution processing, error evaluation, and cycle ambiguity resolution. Finally, it discusses applications and provides a sample of experimental results.

I. Overview

The fundamental principle of attitude determination with GPS and multiple antennas is shown in Fig. 1. The GPS satellite is so distant relative to the antenna separation that arriving wavefronts can be considered as effectively planar. A signal traveling at the speed of light arrives at the antenna closer to the satellite slightly before reaching the other. By measuring the difference in carrier phase between the antennas, a receiver can determine the relative range between the pair of antennas. With the addition of carrier phase measurements from multiple satellites using three or more antennas, the receiver can estimate the full three-axis attitude of an object.

Early experimental work employed TRANSIT satellites for attitude determination.[1] Since then, many GPS receivers have been developed or adapted for carrying out attitude determination. Examples of these include implementations by Magnavox,[2] Trimble,[3] TI, [4] and Ashtech.[5]

In the conventional relative position fix (for example, between two survey receivers), range difference measurements between the two receivers from *four* GPS satellites are required to solve for the three components of Cartesian relative position and receiver clock time bias. For attitude determination, it is possible

*Research Associate, Department of Aeronautics and Astronautics; Manager of GPS Precision Landing.

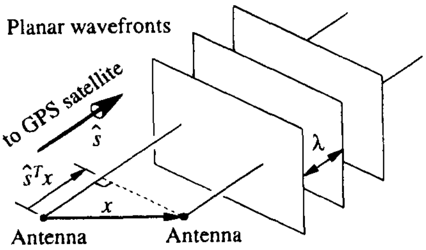


Fig. 1 Attitude geometry.

to configure the receiver so that *only two GPS satellites in view are explicitly required.* There are two reasons for this.

1) *Common Time Reference:* If the receiver is designed as shown in Fig. 2, so that each signal path shares a common time reference, the phase difference measurement precision is maximized. Because any local oscillator variations affect both signal paths identically, these variations cancel out in the final differencing process. Therefore, the measurements are *independent* of the receiver clock bias. Because of the electrical connection between antennas, only signals from *three* GPS satellites are required to find the three Cartesian components of relative antenna position.

2) *Fixed Baseline Configuration:* For attitude determination, the relative mechanical placement of the antennas must be known in advance. Given the additional rigid constraint on relative antenna placement on the vehicle, another satellite measurement can be dropped. Therefore, a minimum of only *two* GPS satellites are required for an attitude fix.

This result provides some very practical benefits. First, overall solution integrity is improved considerably. Because the operational GPS constellation provides at least four satellites in view, attitude solutions are, in general, strongly overdetermined. Occasional cycle slips can be detected and isolated in real time. Second, when the vehicle attitude tips to extremes (such as with an aircraft in a steep angle of bank), attitude solutions are uninterrupted as long as at least two satellites are in view.

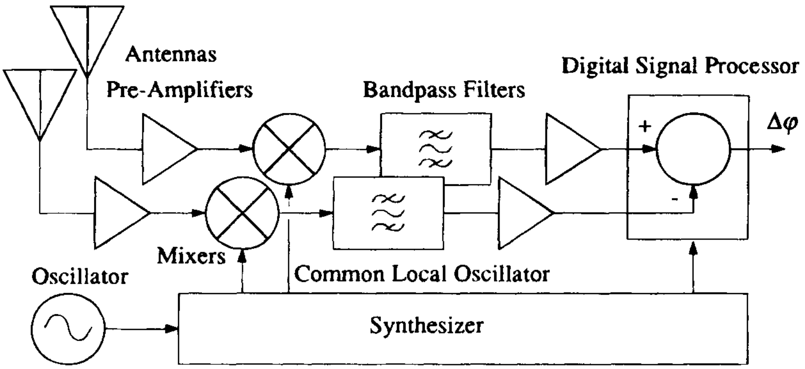


Fig. 2 Common local oscillator.

Referring to Fig. 3, the measured differential phase, $\Delta\varphi$ (measured in wavelengths), is proportional to the projection (vector dot product) of the baseline vector $\boldsymbol{x}$ (3×1), measured in wavelengths (cycles), into the line of sight unit vector to the satellite, $\hat{s}$ (3×1), for baseline i and satellite j. However, as implied in the figure the GPS receiver initially only measures the fractional part of the differential phase. The integer component k must be resolved through independent means before the differential phase measurement can be interpreted as a differential range measurement. The resulting expression is then $\Delta\varphi_{ij} = \hat{s}_j^T\boldsymbol{x}_i - k_{ij} + v_{ij}$, where v_{ij} is additive, time-correlated measurement noise from the relative ranging error sources discussed in Sec. V. Note that in this chapter as a matter of convention, the integer k_{ij} is treated as a *constant* as long as continuous lock is maintained (i.e., until a cycle slip occurs) on that combination of satellite and baseline. In other words, the initial allocation of integer component between k and $\Delta\varphi$ is *arbitrary.* As the satellite–baseline geometry changes with time, it is assumed that the *receiver tracking loops keep automatic track of integer wrap-arounds* in the $\Delta\varphi$ measurements as they occur (i.e., they track the total change in $\Delta\varphi$, including the integer part). Thus, the only ambiguity is the initial value of the integer k. Also note that the line bias attributable to electrical path length differences is not treated in this expression or those which follow, because it can generally be removed through receiver calibration.

II. Fundamental Conventions for Attitude Determination

Although a full discussion of the mathematical tools generally used for attitude determination is beyond the scope of this book, some introductory material on the fundamental conventions for attitude determination is supplied here as a minimum basis for understanding coordinate transformations and attitude parameterization. For a more in-depth description of these concepts, see Ref. 6.

In attitude determination, we typically are concerned with describing a vehicle system in two separate reference frames: the local horizontal (or, alternatively,

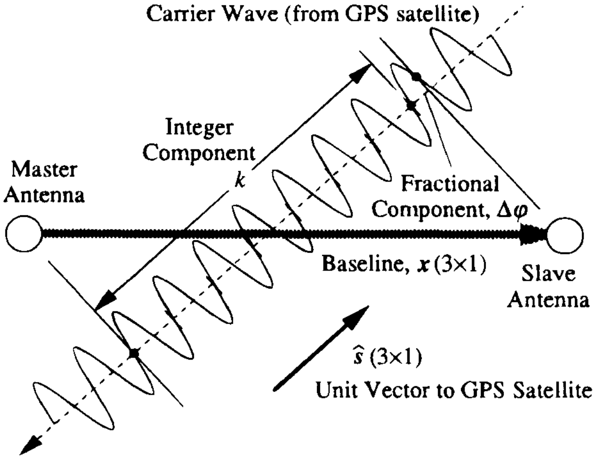

Carrier Wave (from GPS satellite)
Integer
Component
k
Master
Antenna
Fractional
Component, $\Delta\varphi$
Baseline, $\boldsymbol{x}$(3×1)
Slave
Antenna
$\hat{s}$ (3×1)
Unit Vector to GPS Satellite

Fig. 3 Observation geometry.

an inertial) coordinate system and the vehicle body coordinate system. As shown in Fig. 4, the vehicle body coordinate axes (dashed; designated as unit vectors $\boldsymbol{x}', \boldsymbol{y}'$, and $\boldsymbol{z}'$) are generally rotated with respect to the local horizontal coordinate frame axes (solid; designated as unit vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$). Each coordinate frame is right-handed (i.e., $\boldsymbol{x}$ crossed into $\boldsymbol{y}$ equals $\boldsymbol{z}$). A given vector $\boldsymbol{r}$ (3×1) is expressed in the local horizontal coordinate frame. The same vector $\boldsymbol{r}$ can also be expressed as $\boldsymbol{r}'$ in the vehicle body (primed) reference frame through a *coordinate transformation:* $\boldsymbol{r}' = A\boldsymbol{r}$.

The matrix A (3×3) is known as the *attitude matrix* or *direction cosine matrix.* The easiest way to construct the attitude matrix is by assembling the dot products of the orthogonal coordinate frame unit vectors:

$$A = \begin{bmatrix} \boldsymbol{x}'\cdot\boldsymbol{x} & \boldsymbol{x}'\cdot\boldsymbol{y} & \boldsymbol{x}'\cdot\boldsymbol{z} \\ \boldsymbol{y}'\cdot\boldsymbol{x} & \boldsymbol{y}'\cdot\boldsymbol{y} & \boldsymbol{y}'\cdot\boldsymbol{z} \\ \boldsymbol{z}'\cdot\boldsymbol{x} & \boldsymbol{z}'\cdot\boldsymbol{y} & \boldsymbol{z}'\cdot\boldsymbol{z} \end{bmatrix}$$

Although there are nine elements in the matrix, they are not all independent. There are really only three DOF because of the orthonormal constraints ($A^TA = I$) placed on the transformation. The inverse of any attitude matrix is simply its transpose.

The most commonly accepted convention for defining coordinate frames and rotation angles is shown in Fig. 5. Rotations are defined in a specific Euler sequence about the coordinate axes. The axes of the local horizontal frame are such that the $\boldsymbol{x}$ axis points due north, the $\boldsymbol{y}$ axis points due east, and the $\boldsymbol{z}$ axis points directly downward along the local vertical to complete the right-handed set of axes. The body reference frame is fixed to the aircraft so that the $\boldsymbol{x}'$ (roll) axis points out the nose, the $\boldsymbol{y}'$ axis points to the right along the wing, and the $\boldsymbol{z}'$ axis points out the belly to complete the right-handed coordinate axis set. The figure shows the body frame when the pitch and roll angles are zero. For this special case, the pitch axis is aligned with the $\boldsymbol{y}'$ axis, and the heading axis is aligned with the $\boldsymbol{z}'$ axis (local vertical). When the heading, pitch, and roll angles are all zero, the body frame is aligned with the local horizontal frame ($A = I$).

Given this introduction, the three attitude angles (heading, pitch, and roll) may then specify the vehicle attitude. Starting from the reference attitude (where

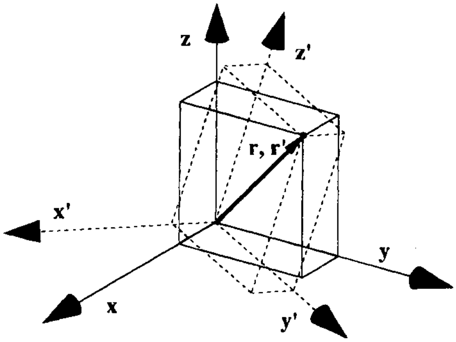


Fig. 4 Coordinate transformation.

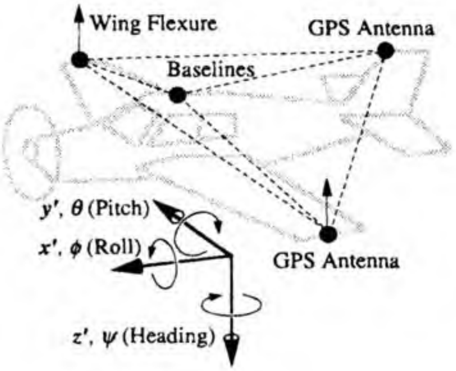


Fig. 5 Aircraft geometry.

the body and local horizontal coordinate frames are aligned), the body frame is rotated (always in a positive, right-handed sense) about the local vertical downward z axis by the heading angle ψ. Then, the body frame is rotated about the new pitch axis by the pitch angle θ. Finally, the body frame is rotated about the roll x' axis by the roll angle ϕ. The resulting attitude matrix A can be shown to be as follows:

$$A = \begin{bmatrix} \cos\theta\cos\psi & \cos\theta\sin\psi & -\sin\theta \\ -\cos\phi\sin\psi + \sin\phi\sin\theta\cos\psi & \cos\phi\cos\psi + \sin\phi\sin\theta\sin\psi & \sin\phi\cos\theta \\ \sin\phi\sin\psi + \cos\phi\sin\theta\cos\psi & -\sin\phi\cos\psi + \cos\phi\sin\theta\sin\psi & \cos\phi\cos\theta \end{bmatrix}$$

III. Solution Processing

This section discusses how differential phase measurements can be converted into attitude solutions. To clarify presentation, it is first assumed that the cycle ambiguities are already known. Discussion of the processes for resolving integers is deferred until Sec. IV. If the integers are known, then the differential *phase* measurements can be treated explicitly as differential *range* measurements through the relationship $\Delta r = \Delta\varphi + k$. Then the process of attitude determination consists of converting these differential range measurements into attitude solutions. An optimal attitude solution for a given set of range measurements Δr_{ij} taken at a single epoch for baseline i and satellite j is obtained by minimizing the quadratic attitude determination cost function:

$$J(A) = \sum_{i=1}^{m} \sum_{j=1}^{n} (\Delta r_{ij} - b_i^T A \hat{s}_j)^2$$

for the m baseline and n satellites, where $\boldsymbol{b}$ (3×1) is the baseline vector defined in the body frame, $\hat{\boldsymbol{s}}$ (3×1) is the line of sight to the GPS satellite given in the local horizontal frame, and A (3×3), the variable to be used in minimization, is the right-handed, orthonormal attitude transformation ($\det A = 1$, $A^T A = I$) from the local horizontal frame to the body frame.

Given a trial attitude matrix A_0, a better estimate may be obtained by linearizing this cost function about the trial solution and solving for a correction matrix δA.

Solving for the best correction matrix during iteration p yields a new and better trial matrix for iteration $p + 1$, so that $A_{p+1} = \delta A_p A_p$. A simple correction matrix can be constructed of small-angle rotations, so that $\delta A(\boldsymbol{\delta\theta}) \cong \mathrm{I} + \Theta^{\times}$, where I (3×3) is the identity matrix, and $\boldsymbol{\delta\theta}$ (3×1) is a vector of small-angle rotations about the following three body frame axes:

$$\boldsymbol{\delta\theta} = \begin{bmatrix} \delta\theta_{x'} \\ \delta\theta_{y'} \\ \delta\theta_{z'} \end{bmatrix}$$

and $\Theta^{\times}$ (3×3) is the skew-symmetric matrix associated with the vector $\boldsymbol{\delta\theta}$.

$$\Theta^{\times} = \begin{bmatrix} 0 & -\delta\theta_{z'} & \delta\theta_{y'} \\ \delta\theta_{z'} & 0 & -\delta\theta_{x'} \\ -\delta\theta_{y'} & \delta\theta_{x'} & 0 \end{bmatrix}$$

so that $\Theta^{\times}\boldsymbol{b} = \boldsymbol{\delta\theta} \times \boldsymbol{b}$. The attitude cost function becomes the following:

$$J(\boldsymbol{\delta\theta})\,|_{A_0} \cong \sum_{i=1}^{m}\sum_{j=1}^{n}[\Delta r_{ij} - \boldsymbol{b}_i^T(I + \Theta^{\times})A_0\hat{\boldsymbol{s}}_j]^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(\delta r_{ij} - \boldsymbol{b}_i^T\Theta^{\times}A_0\hat{\boldsymbol{s}}_j)^2$$

where $\delta r_{ij} \equiv \Delta r_{ij} - \boldsymbol{b}_i^T A_0 \hat{\boldsymbol{s}}_j$. The measurement geometry is described by the rightmost term, which may be rewritten directly in terms of the three attitude correction angles about each axis $\boldsymbol{b}_i^T\Theta^{\times}A_0\hat{\boldsymbol{s}}_j = \hat{\boldsymbol{s}}_j^T A_0^T \Theta^{\times}\boldsymbol{b}_i = \hat{\boldsymbol{s}}_j^T A_0^T B_i^{\times}\boldsymbol{\delta\theta}$.

Because the right-hand side of this result can also be written as

$$[(A_0\hat{\boldsymbol{s}}_j) \times \boldsymbol{b}_i] \bullet \boldsymbol{\delta\theta},$$

the implication is that the attitude angle sensitivity to a measurement from a given baseline and GPS satellite is simply the *cross-product* of the line-of-sight vector with the baseline vector. The linearized cost function may now be written as follows:

$$\delta J(\boldsymbol{\delta\theta})\,|_{A_0} = \|H\boldsymbol{\delta\theta} - \boldsymbol{\delta r}\|_2^2$$

where $\boldsymbol{\delta r}$ is the vector formed by stacking all measurements, and H is the observation matrix formed by stacking the measurement geometry for each separate measurement:

$$H = \begin{bmatrix} \vdots \\ \hat{\boldsymbol{s}}_j^T A_0^T B_i^{\times} \\ \vdots \end{bmatrix}$$

The estimate for A is then refined iteratively until the process converges to the numerical precision of the computer.

In cases where the baseline array is non–co-planar, there is an algorithm for carrying out the attitude calculation approximately an order of magnitude faster. This approach is based on solving "Wahba's Problem" of attitude determination using vector observations.[7]

IV. Cycle Ambiguity Resolution

As suggested in Fig. 3, cycle ambiguity resolution is the process of determining the integer number of wavelengths that lie between a given pair of antennas along a particular line of sight. It is the key initialization step that must be performed before attitude determination using GPS can commence.

A. Baseline Length Constraint

Consider a platform with a single baseline constructed from two antennas. The baseline vector originates at the master antenna and ends at the slave antenna. Because differential positioning is employed, no generality is sacrificed by assuming that the tail of the vector stays fixed in space, as shown in Fig. 6. The possible positions of the slave antenna are constrained to lie on the surface of a virtual sphere of radius equal to the baseline length.

B. Integer Searches

The most brute force method of resolving the integers is the search method. In an integer search, all possible combinations of candidate integers (which can number in the hundreds of millions for antenna separations of even just a few meters) are systematically checked against a cost function until (it is hoped) the correct set is found.

Although search techniques work accurately and quickly for smaller baselines (on the order of several carrier wavelengths), they are vulnerable to erroneous solutions with longer baselines or when few satellites are visible. Although many creative techniques have been synthesized for maximizing the execution speed of the search process,[8–10] searches still occasionally suffer from ambiguous results.

The search technique is depicted to scale in Fig. 7 for a single 4λ baseline, three satellites, and 3σ multipath error. The instantaneous satellite line-of-sight vectors are depicted with arrows. Possible integer values are shown as concentric bands about the line-of-sight vectors. The correct integer set is indicated in white. Any place on the sphere where the concentric bands for all three satellites intersect (indicated in black) is a viable baseline orientation candidate. *Note that at any one instant, there is not a unique solution.*

C. Motion-Based Methods

Although lacking the near-instantaneous start-up time of integer search methods, motion-based methods are unmatched for providing the highest level of

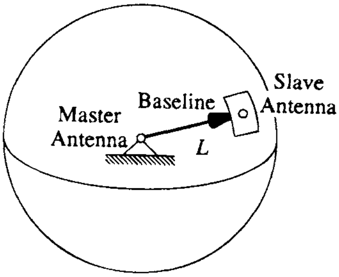


Fig. 6 Length constraint.

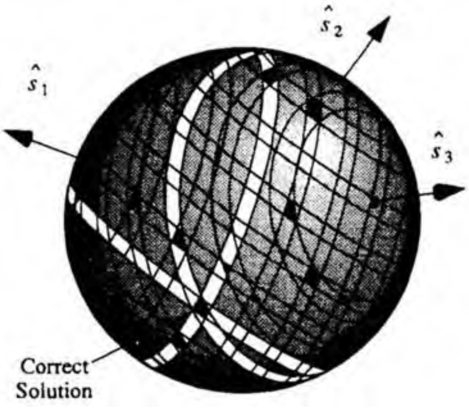


Fig. 7 Integer search.

overall solution integrity.[11] Motion-based integer resolution algorithms make use of the abundance of information provided by platform or GPS satellite motion. This attitude motion modulates the relative carrier phase with a signature that may be used to identify the cycle ambiguities. If the motion occurs rapidly enough, this process is complete within seconds. Rather than constraining cycle ambiguities to lie on integer values, motion-based methods estimate the cycle ambiguities as continuous biases. Checking without imposing the integer constraint that the bias values indeed lie near integer values provides a unique, unambiguous solution with extraordinary integrity—even when there are only a few satellites in view.

For aircraft applications, natural attitude motion consists of banks, turns, or attitude perturbations excited by turbulence. If the baseline vectors are non–co-planar, even a turn on the ground (about a single axis) is sufficient for cycle ambiguity resolution.

Without knowledge of the integers, it is possible to determine the Cartesian position of the slave antenna relative to its unknown starting point. The differential phase measurement equation may be expressed in compact vector and matrix notation for the n satellites in view (neglecting ranging noise):

$$\boldsymbol{\Delta\varphi} = S^T\boldsymbol{x} - \boldsymbol{k}$$

where

$$\boldsymbol{\Delta\varphi}(n \times 1) = \begin{bmatrix} \Delta\varphi_1 \\ \Delta\varphi_2 \\ \vdots \\ \Delta\varphi_n \end{bmatrix}, \qquad S(3 \times n) = [\hat{\boldsymbol{s}}_1 \quad \hat{\boldsymbol{s}}_2 \quad \cdots \quad \hat{\boldsymbol{s}}_n], \ \boldsymbol{k}(n \times 1) = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$

Suppose a baseline moves from Cartesian position $\boldsymbol{x}^{(0)}$ to $\boldsymbol{x}^{(1)}$. The measured change in differential range is given by $\boldsymbol{\Delta\varphi}^{(1)} - \boldsymbol{\Delta\varphi}^{(0)} = S^T[\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}] = S^T\boldsymbol{\Delta x}$. Because the integer ambiguity $\boldsymbol{k}$ cancels out of the above expression, and the satellite line-of-sight vectors are known, we can solve for the displacement vector

$\Delta\boldsymbol{x}$ (3×1), explicitly using a linear least-squares fit. For the integer resolution processing $n = 3$ satellites are used to determine the relative location of the slave antenna undergoing motion. (After cycle resolution is complete, two satellites are required for attitude determination.) It has been assumed that the position displacement is occurring on a much faster time scale than that of the satellite line-of-sight vectors $\hat{\boldsymbol{s}}$.

As baseline motion is occurring, it is possible to accumulate a set of displacement vectors over a short interval of time. If the platform moves by a large angle, the set of displacement vectors can be used to calculate an initial guess to initialize a nonlinear least-squares fit.[4]

A two-dimensional representation of the rigid body antenna mounting constraints is shown in Fig. 8. Suppose that the baseline vector $\boldsymbol{x}$ rotates in space. Here the baseline vector is moved (displaced) by the vector $\Delta\boldsymbol{x}$ to two different locations at two different times, 1 and 2. If a line is constructed perpendicular to each $\Delta\boldsymbol{x}$ displacement vector passing through its midpoint, the center of the circle must be included on that line. By simultaneously considering each $\Delta\boldsymbol{x}$ vector, the center of the circle can be located, along with the initial position $\boldsymbol{x}$ of the baseline.

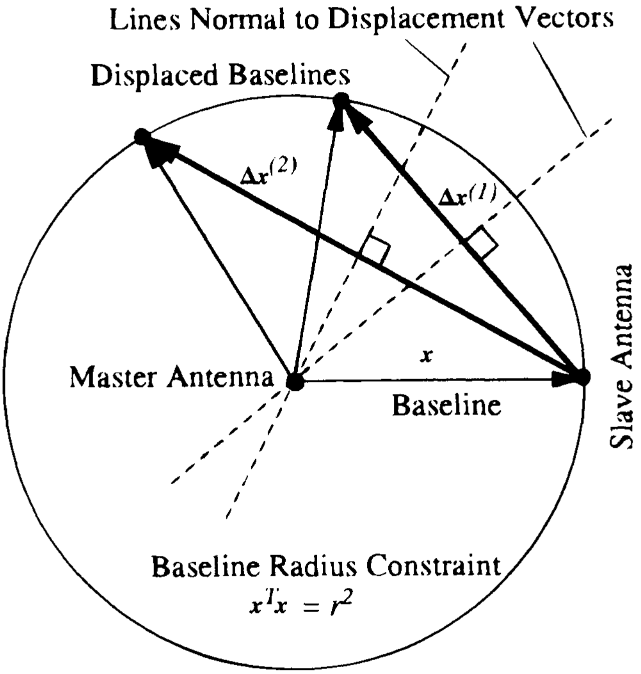


Fig. 8 Large angle motion: initial guess

Mathematically, the solution may be developed by constructing the square of the norm of the rotated baseline vector $\boldsymbol{x} + \Delta\boldsymbol{x}$, given as follows:

$$(\boldsymbol{x} + \Delta\boldsymbol{x})^T(\boldsymbol{x} + \Delta\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x} + 2\Delta\boldsymbol{x}^T\boldsymbol{x} + \Delta\boldsymbol{x}^T\Delta\boldsymbol{x}$$

Noting that the left side is equal to the square of the baseline length, as is $\boldsymbol{x}^T\boldsymbol{x}$, the two terms may be canceled, leaving $2\Delta\boldsymbol{x}^T\boldsymbol{x} = -\Delta\boldsymbol{x}^T\Delta\boldsymbol{x}$.

Different $\Delta\boldsymbol{x}$ vectors taken at N different times (indicated by a superscript in parentheses) may be stacked into matrix form as follows:

$$2\begin{bmatrix} \Delta\boldsymbol{x}^{(1)T} \\ \Delta\boldsymbol{x}^{(2)T} \\ \vdots \\ \Delta\boldsymbol{x}^{(N)T} \end{bmatrix} \boldsymbol{x} = -\begin{bmatrix} \Delta\boldsymbol{x}^{(1)T}\Delta\boldsymbol{x}^{(1)} \\ \Delta\boldsymbol{x}^{(2)T}\Delta\boldsymbol{x}^{(2)} \\ \vdots \\ \Delta\boldsymbol{x}^{(N)T}\Delta\boldsymbol{x}^{(N)} \end{bmatrix}$$

Then the baseline solution x may be obtained through a linear least-squares fit. Each $\Delta\boldsymbol{x}$ vector defines a subspace in which the slave antenna must lie. The baseline is then the point that comes closest to this condition in a least-squares sense. A convenient shorthand notation for this least-squares fit is given by $2\Delta\boldsymbol{X}^T\boldsymbol{x} = -\text{diag}(\Delta\boldsymbol{X}^T\Delta\boldsymbol{X})$, $\Delta\boldsymbol{X}$ $(3 \times N) \equiv [\Delta\boldsymbol{x}^{(1)} \quad \Delta\boldsymbol{x}^{(2)} \quad \cdots \quad \Delta\boldsymbol{x}^{(N)}]$ and diag(•) defines a vector $(N \times 1)$ comprised of the diagonal elements of the argument matrix $(N \times N)$.

Unfortunately, for the single baseline case, large-angle rotation about the two axes perpendicular to the baseline is always required to resolve completely the integer ambiguities using motion. To avoid the shortcoming of requiring two-axis motion perpendicular to each baseline, information from multiple baselines can be combined into a single simultaneous estimation equation.[12] The constraint equation between different baselines i and j yields $(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i)^T(\boldsymbol{x}_j + \Delta\boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j + \boldsymbol{x}_i^T\Delta\boldsymbol{x}_j + \Delta\boldsymbol{x}_i^T\boldsymbol{x}_j + \Delta\boldsymbol{x}_i^T\Delta\boldsymbol{x}_j$.

Again, the dot product of each baseline pair is constant; hence, the corresponding term may be canceled from both sides of the equation, leaving $\Delta\boldsymbol{x}_j^T\boldsymbol{x}_i + \Delta\boldsymbol{x}_i^T\boldsymbol{x}_j = -\Delta\boldsymbol{x}_i^T\Delta\boldsymbol{x}_j$.

Combining $\Delta\boldsymbol{x}$ measurements from N different times (typically 10–30 epochs), this form may be expanded as follows:

$$\begin{bmatrix} \Delta\boldsymbol{x}_j^{(1)T} \\ \Delta\boldsymbol{x}_j^{(2)T} \\ \vdots \\ \Delta\boldsymbol{x}_j^{(N)T} \end{bmatrix} \boldsymbol{x}_i + \begin{bmatrix} \Delta\boldsymbol{x}_i^{(1)T} \\ \Delta\boldsymbol{x}_i^{(2)T} \\ \vdots \\ \Delta\boldsymbol{x}_i^{(N)T} \end{bmatrix} \boldsymbol{x}_j = -\begin{bmatrix} \Delta\boldsymbol{x}_i^{(1)T}\Delta\boldsymbol{x}_j^{(1)} \\ \Delta\boldsymbol{x}_i^{(2)T}\Delta\boldsymbol{x}_j^{(2)} \\ \vdots \\ \Delta\boldsymbol{x}_i^{(N)T}\Delta\boldsymbol{x}_j^{(N)} \end{bmatrix}$$

Invoking the same matrix notation as above, the entire initial guess for the case of the three baselines shown in Fig. 9 can be combined into a single, unified least-squares fit equation:

$$\begin{bmatrix} \Delta \boldsymbol{X}_2^T & \Delta \boldsymbol{X}_1^T & 0 \\ \Delta \boldsymbol{X}_3^T & 0 & \Delta \boldsymbol{X}_1^T \\ 0 & \Delta \boldsymbol{X}_3^T & \Delta \boldsymbol{X}_2^T \\ 2\Delta \boldsymbol{X}_1^T & 0 & 0 \\ 0 & 2\Delta \boldsymbol{X}_2^T & 0 \\ 0 & 0 & 2\Delta \boldsymbol{X}_3^T \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \end{bmatrix} = - \begin{bmatrix} \mathrm{diag}(\Delta \boldsymbol{X}_1^T \Delta \boldsymbol{X}_2) \\ \mathrm{diag}(\Delta \boldsymbol{X}_1^T \Delta \boldsymbol{X}_3) \\ \mathrm{diag}(\Delta \boldsymbol{X}_3^T \Delta \boldsymbol{X}_2) \\ \mathrm{diag}(\Delta \boldsymbol{X}_1^T \Delta \boldsymbol{X}_1) \\ \mathrm{diag}(\Delta \boldsymbol{X}_2^T \Delta \boldsymbol{X}_2) \\ \mathrm{diag}(\Delta \boldsymbol{X}_3^T \Delta \boldsymbol{X}_3) \end{bmatrix}$$

The left-hand matrix is now $6N \times 9$, the solution vector is 9×1, and the right-hand vector is $6N \times 1$. This same matrix structure applies to any number of baselines. For the case of three or more baselines, two important advantages fall out of this approach.

First, with motion about any two *arbitrary* axes, no a priori information about antenna placement is required to unambiguously solve for all three of the three initial baseline vectors (three components each, nine total dimensions). It is the measurements themselves that are providing all the geometrical information. Therefore, this approach could be adapted to perform *in situ* self-calibration of GPS baselines during normal operation.

Second, by incorporating the known baseline constraints in the case of non–co-planar baseline configurations, *motion about a single axis of rotation is entirely adequate for an unambiguous baseline vector solution.*

1. Measurement Refinement

To refine the initial guess iteratively, a new cost function (modeled after the one defined in Sec. III) is employed:

$$J(A^{(1)}, A^{(2)}, \ldots, A^{(N)}, \boldsymbol{k}) = \sum_{\ell=1}^{N} \sum_{i=1}^{m} \sum_{j=1}^{n} (\Delta\varphi_{ij}^{(\ell)} + \boldsymbol{k}_{ij} - \boldsymbol{b}_i^T A^{(\ell)} \hat{\boldsymbol{s}}_j^{(\ell)})^2$$

where $A^{(\ell)}$ is the attitude matrix at each epoch ℓ, and $\boldsymbol{k}$ is a vector of the cycle ambiguities for all the baseline and antenna combinations. The ambiguities are estimated as continuous variables.

The problem is to find the independent attitude matrices at each epoch and the set of integers (which applies to all epochs) that minimize the stated cost function. For each epoch, it is possible to convert the initial guess of baseline position given above into an initial guess for the attitude A_0. As shown in Sec. III, the cost function can then be linearized about this initial guess. The attitude component of the state variables to be estimated consists of perturbations in the vehicle attitude about all three axes for every epoch l under consideration. The

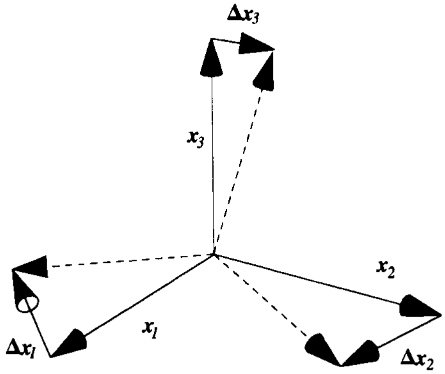


Fig. 9 Non–co-planar baseline rotation.

resulting linearized equations are as follows:

$$\left[\begin{array}{cccc|c} H_1 & 0 & 0 & 0 & -I \\ 0 & H_2 & & 0 & -I \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & H_N & -I \end{array}\right] \left[\begin{array}{c} \delta\boldsymbol{\theta}^{(1)} \\ \delta\boldsymbol{\theta}^{(2)} \\ \vdots \\ \delta\boldsymbol{\theta}^{(N)} \\ \hline \boldsymbol{k} \end{array}\right] = \left[\begin{array}{c} \delta\boldsymbol{\varphi}^{(1)} \\ \delta\boldsymbol{\varphi}^{(2)} \\ \vdots \\ \delta\boldsymbol{\varphi}^{(N)} \end{array}\right]$$

where each $H_l(mn \times 3)$ is the sensitivity matrix of changes in measured differential phase with respect to rotations about each of the three axes of attitude, and I $(mn \times mn)$ is the identity matrix. As was shown in Sec. III, each row of H is given by $(A_0^{(l)} S_j^{(l)}) \times \boldsymbol{b}_i$. The other state variables are as follows:

$$\delta\boldsymbol{\theta}^{(\ell)}(3 \times 1) = \begin{bmatrix} \delta\theta_x^{(\ell)} \\ \delta\theta_y^{(\ell)} \\ \delta\theta_z^{(\ell)} \end{bmatrix} \quad \text{and} \quad \boldsymbol{k}(mn \times 1) = \begin{bmatrix} k_{11} \\ k_{12} \\ \vdots \\ k_{mn} \end{bmatrix}$$

which are vectors of small-angle, three-axis attitude rotation corrections for each time sample l, and a vector of integer biases, respectively, for all combinations of m baselines and n GPS satellites. The right-hand side of the matrix equation is a vector of differential phase residuals, so that the following results:

$$\delta\boldsymbol{\varphi}^{(\ell)}(mn \times 1) = \begin{bmatrix} \delta\varphi_{11}^{(\ell)} \\ \delta\varphi_{12}^{(\ell)} \\ \vdots \\ \delta\varphi_{mn}^{(\ell)} \end{bmatrix}$$

where

$$\delta\varphi_{ij}^{(\ell)} = \Delta\varphi_{ij}^{(\ell)} - \boldsymbol{b}_i^T A_0^{(\ell)} \hat{\boldsymbol{s}}_j^{(\ell)}$$

and $A_0^{(\ell)}$ is the current best estimate of the platform attitude for epoch ℓ.

To this point, it has been assumed that the time scale of platform motion is very much faster than that of the GPS satellites. It is also possible for the time scales of motion for the two to be comparable, such as in space or marine applications. That *quasi static* case can be treated by applying exactly the same nonlinear, least-squares fit equations. The principal difference is that the time interval of measurement collection is increased.

2. *Static Integer Resolution*

In the static case, the solution may also be refined iteratively by linearizing the observation equation. However, because the static platform attitude is the same for all epochs l, the form of linearized observation equation is as follows:

$$\begin{bmatrix} H_1 & | & -I \\ H_2 & | & -I \\ \vdots & | & \vdots \\ H_N & | & -I \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta\theta} \\ \hline \boldsymbol{k} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\delta\varphi}^{(1)} \\ \boldsymbol{\delta\varphi}^{(2)} \\ \vdots \\ \boldsymbol{\delta\varphi}^{(N)} \end{bmatrix}$$

Again, the right-hand side is a vector of differential phase residuals. The final solution for the vector k yields the integer ambiguities. For non–co-planar baselines, the estimation process usually has enough information to resolve the ambiguities reliably after about 10 min of satellite motion—even with only two satellites in view.

D. Alternative Means for Cycle Ambiguity Resolution

Although motion-based methods for cycle ambiguity resolution are certainly not the only means for system initialization, they undoubtedly have the highest integrity—especially if no external information is available. There are at least two other "instantaneous" approaches that may also be employed with the potential disadvantage that they require more hardware to implement:

1) *Multiple GPS Antennas:* By using small and large baselines together, it is possible to resolve cycle ambiguities in an explicit sequence by starting with the small baselines (where there is little ambiguity) and working one's way out to the larger baselines. Although the entire process is rapid, additional antennas are required.

2) *Multiple GPS Observables:* Another alternative is to use code ranging to establish the dual-frequency carrier, wide-lane ambiguity, allowing the L_1 cycle ambiguities to be resolved. However, on rare occasions the method may still fail to establish the correct wide-lane ambiguity, and a more expensive dual-frequency receiver is required.

V. Performance

This section examines key aspects of the overall performance of attitude determination using GPS and quantifies the most significant sources of error in attitude determination.

A. Geometrical Dilution of Precision for Attitude

The H matrix from Sec. III is the best means for evaluating the attitude fix accuracy. In general, the attitude error is a function of the satellite geometry, baseline geometry, *and* instantaneous vehicle attitude. Given a differential ranging error of σ (typically 5mm), an estimate of the attitude covariance matrix P is given by $P = (H^T H)^{-1}\sigma^2$, where the 1 σ pointing error (in radians) for any given attitude axis is given by the square root of the corresponding diagonal element of this 3 × 3 covariance matrix. The diagonal elements correspond to small rotations about the $\boldsymbol{x}'$, $\boldsymbol{y}'$, and z' body frame coordinate axes, respectively.

For generality, the baseline and satellite line-of-sight vectors can be concatenated into matrices B $(3 \times m)$ and S $(3 \times n)$, where $B = [\boldsymbol{b}_1 \quad \boldsymbol{b}_2 \quad \cdots \quad \boldsymbol{b}_m]$ and $S = [\hat{\boldsymbol{s}}_1 \quad \hat{\boldsymbol{s}}_2 \quad \cdots \quad \hat{\boldsymbol{s}}_n]$.

For the ideal baseline configuration where $BB^T = L^2 I$ (where L is the effective baseline length), I is the 3 × 3 identity matrix, and each of the n GPS satellites is in view of all the antennas on the vehicle, it can be shown that the attitude covariance simplifies to

$$P = [nI - ASS^T A^T]^{-1}\left(\frac{\sigma}{L}\right)^2$$

This form suggests a convenient means for characterizing the suitability of the constellation geometry for attitude determination. As an analog to GDOP, ADOP, the geometric dilution of precision for attitude, is defined by considering the geometrical component of the covariance matrix. Invoking the invariance of the matrix trace with respect to coordinate rotations, the resulting total angular pointing error σ_θ may be written as follows:

$$\sigma_\theta = (\text{ADOP})\,\frac{\sigma}{L}$$

where

$$\text{ADOP} \equiv \sqrt{\text{trace}[(nI - SS^T)^{-1}]}$$

The quantity ADOP is defined even when there are only two satellites in view, the minimum number required for three-axis attitude determination. Its value is generally around unity or smaller, indicating that the GPS constellation consistently provides a favorable geometry for attitude determination.

Thus, a further approximation for the attitude error can be made by simply neglecting the satellite geometry term (ADOP) and considering it to be near unity, so $\sigma_\theta \cong \sigma/L$.

The remaining issue is determining what to use for the value of σ. Table 1 offers typical numbers for relative positioning. In all but the highest regimes of dynamics, the largest error source is multipath.

B. Multipath

Multipath is without question the largest source of error in attitude determination using GPS. Although the actual error it produces is highly deterministic (i.e., a function of the specific environment, materials, antenna gain pattern,

Table 1 Attitude determination ranging error sources (1σ)

Sources	Range error m
Multipath (differential range error for a given pair)	~5 mm
Structural distortion (flexure, thermal expansion)	Application-specific
Troposphere	Modelable
Carrier-to-noise ratio	<1 mm
Receiver-specific errors (crosstalk, line bias, interchannel bias)	<1 mm
Total rss differential ranging error (1σ), excluding distortion	~5 mm

geometry, and other factors), practical experience suggests the approximate rule of thumb that the differential ranging error between a pair of hemispherical microstrip patch antennas is about 5 mm, 1 σ.

In most cases, the most practical and cost-effective approach to systems engineering is simply to use GPS attitude determination in those applications for which this standard multipath error of 5 mm would be acceptable. In cases where more accuracy is required, a number of techniques have been proposed for improving multipath errors. A partial list of techniques includes multipath calibration or antenna pattern shaping,[11] inertial aiding,[13] and mathematical multipath modeling.[14] However, such performance enhancements also carry a penalty of cost or complexity.

C. Structural Distortion

Structural distortion (caused by thermal or flexural bending) can be an issue in certain applications. In most cases, it can either be neglected, modeled, or estimated. (An example of estimating wing flexure on an airplane is given in Fig. 12.)

D. Troposphere

The troposphere can often be a source of error in attitude determination. The simplified model in Fig. 10 shows ray propagation from the vacuum of space down toward the Earth's surface. Refraction of the GPS ranging signal causes the ray to bend downward as atmospheric density (and index of refraction) increases. The simplified slab model depicted in Fig. 10 treats the atmosphere as a block of uniform density and index of refraction n_2. Using Snell's law of refraction

$$\frac{\sin \theta_2}{\sin \theta_1} = \frac{n_1}{n_2}$$

where n_1 is unity, and n_2 (depending upon the atmospheric and water vapor

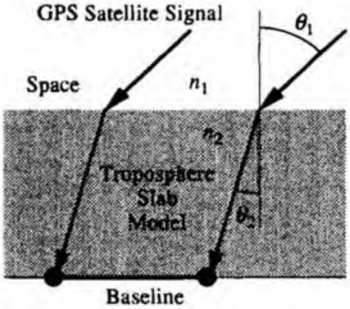


Fig. 10 Atmospheric refraction.

conditions) is somewhere around 1.00026. This simple model can be used to adjust the apparent line-of-sight vectors of the GPS satellites to account for the troposphere in attitude determination applications.

E. Signal-to-Noise Ratio

In applications where it is desirable to track higher dynamics, the tracking loop bandwidth can be opened up (within limitations). The noise on the reconstructed carrier, which dictates the differential range measurement error, is given by the following white noise equation:

$$\sigma = \sqrt{\frac{f_N}{C/N_0}}\,\frac{\lambda}{2\pi}$$

where f_N is the noise bandwidth of the carrier tracking loop, and C/N_0 is the carrier to noise ratio.[15] Typically, this error is dominated by multipath and is smaller than a millimeter for typical tracking parameters ($C/N_0 = 40$ dB-Hz, $f_N = 10$ Hz).

F. Receiver-Specific Errors

Receiver-specific errors, including crosstalk, line bias, and interchannel bias, can be significant sources of error if they are not treated appropriately in a receiver design. Crosstalk between the radio frequency paths for each antenna is an issue, because there is often more than 100 dB of gain along each signal path. Line bias is the nearly constant offset in phase from one antenna to another. A function of both cable length and temperature, line bias is all that remains of the relative clock offset in the design of the common local oscillator. Finally, interchannel bias results from using different hardware channels to measure the carrier phase for each satellite. State-of-the-art receivers employ special techniques to render the errors from all of these effects much smaller than those from multipath.

G. Total Error

Because multipath usually dominates all other error sources (in the absence of significant structural distortion), an approximate and general rule of thumb for attitude determination angular accuracy (in radians) for a representative baseline length of L (in cm) is simply as follows:

$$\boxed{\sigma_\theta \text{ (in radians)} \cong \frac{0.5 \text{ cm}}{L \text{ (in cm)}}}$$

VI. Applications

The capability to use GPS for attitude determination opens up a new realm of applications and opportunities. In the future, it is very likely that the integration of attitude determination into larger systems—including those that use carrier phase for positioning as well—will play a key role in realizing the full potential of GPS. Applications in aviation, spacecraft, and marine areas are summarized below.

A. Aviation

In aviation, heading and attitude sensing using GPS provides a readout that is completely immune to drift and magnetic variation. Many researchers have carried out aircraft experiments to test attitude determination using GPS.[16–19] Figure 11 shows the agreement in roll attitude between GPS and an inertial navigation unit (INU) at a 10-Hz output rate. This flight experiment employed the GPS attitude system that was developed by Stanford University and built

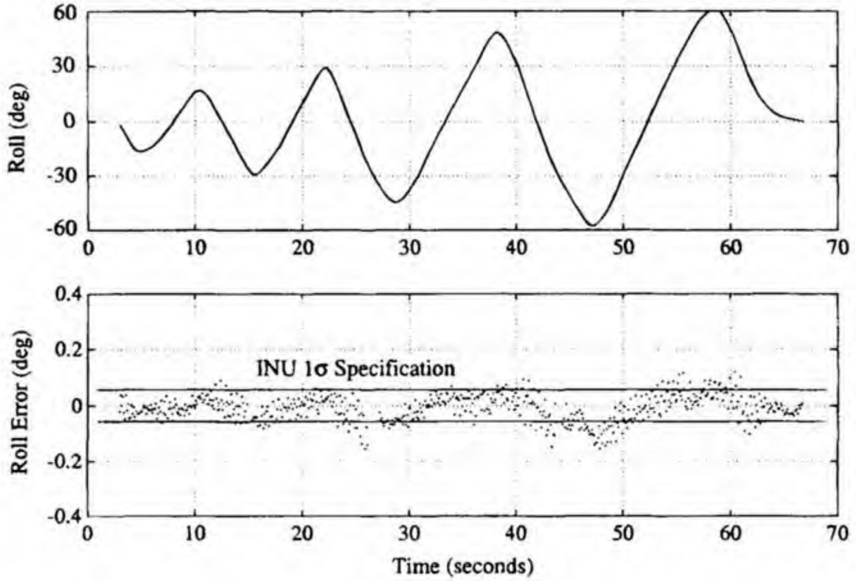


Fig. 11 King Air roll reversals and the inertial navigation unit–GPS agreement.

around a Trimble TANS Quadrex receiver. The system was flown on a NASA Ames King Air, carrying an INU specified to a one sigma accuracy of 0.05 deg. The antennas were mounted on the fuselage, wing tips, and tail, giving the roll component a 16-m baseline. As the aircraft performs roll reversals up to a 60-deg angle of bank, the disparity between roll attitude measured by the two independent sensors seldom exceeds the INU specification. At the steepest angles of bank, the airframe is blocking many of the satellites in view. Sometimes tracking as few as two GPS satellites, the system flawlessly hands off the integers in real time. *In a time span of just a few seconds, the system is tracking a completely new set of satellites with a completely new set of integers.*

A new application of GPS attitude determination is identification of the aircraft dynamic model. The pilot can supply inputs to the controls that excite a dynamic response. The GPS sensor then measures this response to reconstruct an accurate model of the aircraft dynamics. The model then serves as the foundation for optimal autopilot synthesis. Figure 12 shows an example of the characteristic pitch response of an aircraft to a stick input.[12] The "phugoid" response of the aircraft reveals the natural frequency and damping of this mode. It is possible that such model estimation can be carried out continuously in flight, providing a new set of constraints to the position fixes performed by the navigation equipment and, thus, a means for additional integrity checking.

Figure 12 also shows how instantaneous wing flexure is measured to a precision of 1.4-mm rms. The error was evaluated with respect to a best-fit second-order response. Using the same GPS attitude determination system and the antenna arrangement shown in Fig. 5, the structural deformation of the airframe can be used as an indirect means of measuring vertical acceleration.

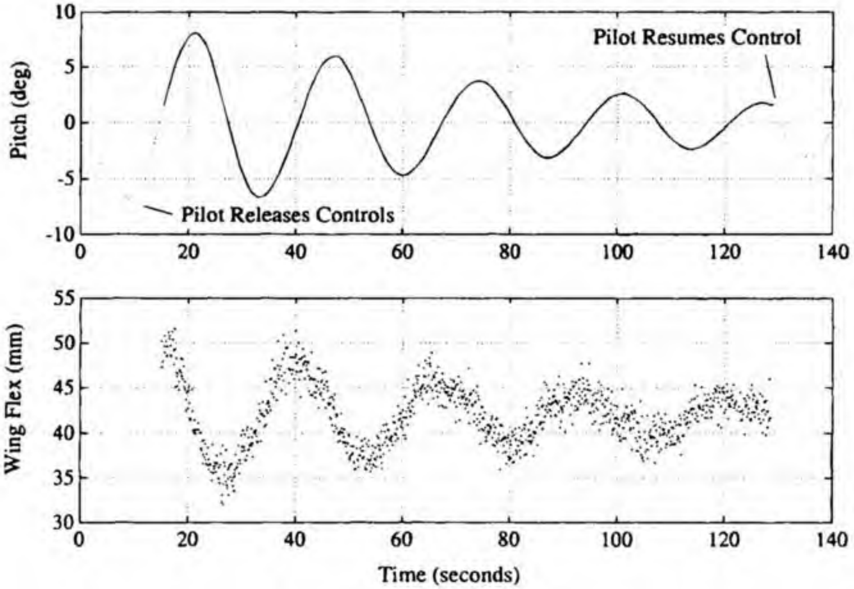


Fig. 12 Piper Dakota phugoid mode state estimates.

By integrating attitude with the enroute navigation, precision landing, collision avoidance, automatic dependent surveillance functions, *a single GPS sensor has the potential to perform the functions of a significant fraction of cockpit instruments currently in use.*

B. Spacecraft

The state-of-the-art in attitude receivers is small (1300 cc), light (~1.5 kg), and low power (~3.5 W), so that one can be carried on just about any spacecraft. Initial experiments with attitude determination on spacecraft[20] indicate that for many types of missions, the GPS may offer significant cost savings. See Chapter 16, this volume, for more information on closed-loop space applications.

C. Marine

In the marine area, the standard for comparison is the gyrocompass. The GPS offers low-cost heading indicator output with rapid start-up times and all-latitude operation. Some of the marine work in attitude determination is described in Refs. 3 and 21.

Attitude determination also provides the potential to point antennas or other directional devices (such as weaponry) on ship-based or other moving platforms. Applying closed-loop control stabilizes the platform against changes in the vehicle orientation.

This partial list of applications hardly begins to address the ultimate potential of attitude determination using the GPS. Given trends in lowering cost, size, weight, and power of GPS technology, it is not inconceivable that backpackers could carry a hand-held portable direction finder that complements the GPS positioning capability. With attitude capability in such a small package, many more applications will undoubtedly arise.

References

[1] Albertine, J. R., "An Azimuth Determination System Utilizing the Navy Navigation Satellites," *Navigation* Vol. 21, No. 1, 1974.

[2] Joseph, K. M., and Deem, P. S., "Precision Orientation: A New GPS Application," International Telemetering Conference, San Diego, CA, Oct., 1983.

[3] Kruczynski, L. R., Li, P. C., Evans, A. G., and Hermann, B. R., "Using GPS to Determine Vehicle Attitude: U.S.S. Yorktown Test Results," *Proceedings of ION GPS-89* (Colorado Springs, CO), Institute of Navigation, Washington, DC, Sept. 1989.

[4] Brown, R., and Ward, P., "A GPS Receiver with Built-in Precision Pointing Capability," *Proceedings, IEEE PLANS, 90, IEEE Position, Location, and Navigation Symposium,* (Las Vegas, NV), Institute of Electrical and Electronics Engineers, New York, March 1990, pp. 83–93.

[5] Kuhl, M., Qin, X., and Cotrell, W., "Design Considerations and Operational Results of an Attitude Determination Unit," *Proceedings of ION GPS-94* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[6] Wertz, J. R. (ed.), *Spacecraft Attitude Determination and Control,* Reidel, Boston, MA, 1985.

[7] Cohen, C. E., Cobb, H. S., and Parkinson, B. W., "Two Studies of High Performance Attitude Determination Using GPS: Generalizing Wahba's Problem for High Output Rates and Evaluation of Static Accuracy Using a Theodolite," *Proceedings of ION GPS-92* (Albuquerque, NM), Institute of Navigation, Washington, DC, Sept. 16–18, 1992.

[8] Hatch, R., "Instantaneous Ambiguity Resolution," KIS Symposium 1990, Banff, Canada, Sept. 1990.

[9] Knight, D., "A New Method of Instantaneous Ambiguity Resolution," *Proceedings of ION GPS-94* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[10] Brown, R. A., "Instantaneous GPS Attitude Determination," *IEEE Aerospace and Electronics Magazine,* June 1992, p. 3.

[11] Cohen, C. E., "Attitude Determination Using GPS," Ph.D. Dissertation, Stanford Univ., Stanford, CA, Dec. 1992.

[12] Cohen, C. E., and Parkinson, B. W., "Aircraft Applications of GPS-Based Attitude Determination: Test Flights on a Piper Dakota," *Proceedings of ION GPS-92* (Albuquerque, NM), Institute of Navigation, Washington, DC, Sept. 16–18, 1992.

[13] Braasch, M., and van Graas, F., "Guidance Accuracy Considerations for Real-Time Interferometric Attitude Determination," *Proceedings of ION GPS-91* (Albuquerque, NM), Institute of Navigation, Washington, DC, Sept. 1991.

[14] Axelrad, P., Comp, C., and MacDoran, P., "Use of Signal-to-Noise Ratio for Multipath Error Correction in GPS Differential Phase Measurements: Methodology and Experimental Results," *Proceedings of ION GPS-94,* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[15] Rath, J., and Ward, P., "Attitude Estimation using GPS," National Technical Meeting ION, San Mateo, CA, Jan. 1989.

[16] Purcell, G. H., Jr., Srinivasan, J. M., Young, L. E., DiNardo, S. J., Hushbeck, E. L., Jr., Meehan, T. K., Munson, T. N., and Yunck, T. P., "Measurement of Aircraft Position, Velocity and Attitude using Rogue GPS Receivers," Fifth International Geodetic Symposium on Satellite Positioning, Las Cruces, NM, March 1989.

[17] van Graas, F., and Braasch, M., "GPS Interferometric Attitude and Heading Determination: Initial Flight Test Results," *Navigation,* Vol. 38, Fall 1991, pp. 297–316.

[18] Cohen, C. E., McNally, B. D., and Parkinson, B. W., "Flight Tests of Attitude Determination Using GPS Compared Against an Inertial Navigation Unit," *Navigation,* Vol. 41, No. 1, 1994.

[19] Cannon, M. E., Sun, H., Owen, T. E., and Meindl, M. A., "Assessment of a Non-Dedicated GPS Receiver System for Precise Airborne Attitude Determination," *Proceedings of ION GPS-94* (Salt Lake City, UT), Institute of Navigation, Washington, DC, Sept. 20–24, 1994.

[20] Cohen, C. E., Lightsey, E. G., Feess, W. A., Parkinson, B. W., "Space Flight Tests of Attitude Determination Using GPS," *International Journal of Satellite Communications,* Vol. 12, Sept.–Oct. 1994, pp. 427–433.

[21] Lu, G., Cannon, M. E., Lachapelle, G., and Kielland, P., "Attitude Determination in a Survey Launch Using Multi-Antenna GPS Technologies," ION National Technical Meeting, San Francisco, CA, Jan. 1993.

Chapter 20

Geodesy

Kristine M. Larson*
University of Colorado, Boulder, Colorado 80309

I. Introduction

GEODESY is the discipline devoted to the measurement of the shape of the Earth and its gravity field in three-dimensional space and time. An ideal geodetic system would provide absolute coordinates of points on the Earth at whatever temporal spacing is required by the geodesist. Until space geodetic techniques were developed, geodetic measurements only indirectly measured changes in the Earth's shape. For example, classical geodetic systems, such as triangulation and trilateration, do not measure position, and thus have significant limitations. In order to resolve changes in the shape of the Earth, the individual angle or length observations must be combined into a network. The final network estimate is formed from an adjustment of the observations, so that the network "closes."[52] Obviously these measurement systems require intervisibility of the observing geodetic stations, making the measurements local in scale (1–50 km). Even at these distances, systematic errors in both triangulation and trilateration grow rapidly with increasing baseline length. Additionally, rotation and translation of the network are indeterminate. In order to interpret the deformation of the network over time, a strain analysis is done. Although useful for interpretation for geodetic data in active tectonic regions, strain analysis is limited by the assumptions of uniformity over the whole network.

Space geodetic techniques such as VLBI (very long baseline interferometry) and SLR (satellite laser ranging) are both more flexible and precise. Visibility requirements for VLBI and SLR are skyward rather than to other observing sites, so that global scale measurements are feasible. Both VLBI and SLR observables can be analyzed to determine three-dimensional station positions. The major drawback of both systems is cost. The VLBI and SLR systems cost millions of dollars to build and millions to maintain and operate. Specially trained personnel

*Assistant Professor, Department of Aerospace Engineering Sciences.

are required to operate the systems, and mobility is limited by their size (multiple vans) and weight (several tons).

The Global Positioning System (GPS) has revolutionized geodesy through its great accuracy, convenience, and global availability. It is fully three-dimensional, and translation and network rotation can be determined, as long as common sites are observed and a stable reference frame is used. With a full GPS constellation, satellite signals, and thus positioning ability, are globally available 24 h a day. The cost of GPS geodesy is several orders of magnitude less than VLBI and SLR. A high-quality GPS receiver costs ~$25,000 and can be installed at a fixed site with minimal cost. GPS data are then easily and inexpensively collected continuously without requiring human intervention. For specific geodetic experiments, GPS receivers can be easily deployed, with the equipment weighing ~30 lb. With proper modeling of the GPS observables, the shape and surface dynamics of the Earth can be unambiguously determined with great precision and accuracy.

In this chapter we concentrate on reviewing GPS geodetic positioning work that has already been published in the refereed literature. The GPS contribution to the measurement of the gravity field is just now being realized, with encouraging initial results from the TOPEX/Poseidon mission,[47] but is not discussed here. First, we briefly summarize the error sources and discuss reference frames issues of interest for precise geodetic analysis with GPS. Following that, we discuss recent geodetic results, focusing in particular on measuring motions of the Earth's crust and its rotation axis.

II. Modeling of Observables

The GPS observables are corrupted by numerous errors, as is discussed in detail in other sections of this book. For high-quality receivers, the phase measurement error is no worse than 1% of the carrier wavelength, or 2 mm. The P-code pseudorange measurement error is several orders of magnitude worse. Both observables, which are modeled as the distance plus the time offset between the satellite and receiver, are corrupted by satellite and receiver instrumental delays and path errors associated with the atmosphere. An accurate receiver location can be determined only with equivalently accurate satellite positions. Because these error sources are discussed in other sections of this book, we simply summarize techniques that are commonly used in precise geodetic softwares.

Although not technically an "error" source, the treatment of "cycle slips" is an important issue in precise geodetic applications. These breaks in the carrier phase data are caused by the receivers or by obstructions in the path of the signal. A reliable algorithm is required to identify and, if possible, repair such slips. Blewitt[4] developed an algorithm that uses the P-code pseudorange to repair slips for undifferenced data. An alternative algorithm was developed by Freymueller[21] to remove cycle slips for receivers without P code. Most software repairs cycle slips in the double differenced data.[25]

Clock drifts, for both satellites and receivers, can be estimated directly at each epoch,[32] or removed via differencing of the data.[25] As discussed in Chapter 16, this volume, direct estimation of the clocks yields greater flexibility and more independent measurements. One limitation of estimating the clocks is that cycle slips must be repaired in the undifferenced data, which can be difficult with codeless receivers when selective availability is active. For determination of relative positions, clock errors are not a limiting error source. Selective availability (SA), which to the GPS analyst is equivalent to a noisy clock, does not have a significant effect on relative positioning accuracy.

The most significant path errors for precise GPS geodetic applications are associated with the atmosphere, which are discussed in the companion volume. The first-order ionospheric errors are eliminated using an appropriate linear combination of the *L*1 and *L*2 phase data. The troposphere has been a more troublesome error source. The component most affected by the tropospheric error is the vertical component, and to a lesser extent the east component.[51] Although it has been suggested that water vapor radiometers (WVRs) could calibrate the tropospheric error, there has been little success in using WVR data in precise geodetic applications without also estimating a parameter representing the zenith troposphere delay. Recently Ware et al.[54] suggested that a WVR can be used to model the wet zenith delay after proper calibration of the WVR. These lengthy calibration procedures would seem to preclude the use of WVRs in general surveying, but they may prove useful at permanent global tracking sites.

Although it is well known that the wet troposphere delay varies appreciably over the course of typical GPS experiments, many of the original studies of GPS precision estimated only one troposphere zenith delay parameter per site per observing session (e.g., Refs. 2 and 17). Most software has since been modified to allow estimation of a piecewise constant troposphere delay. Lichten and Border[32] showed that the wet troposphere zenith delay is accurately characterized as a time-varying parameter with the statistical properties of a random walk.

The remaining path error, multipathing of the GPS signals off objects near the receiver, seems slight for carrier phase data. Multipathing for the pseudorange data is a more serious problem. Although it has been suggested that site-dependent multipath filters could be developed,[12,22] it would seem that significant reduction of multipath will be attributed to improvements in antenna design, receiver signal processing, and better site selection.

The phase center of a GPS antenna varies as a function of the elevation and azimuth angles to the satellite. The magnitude of this variation ranges from several millimeters to several centimeters. To date, most GPS analysis softwares have used a mean GPS phase center to model the phase observables. This was convenient when all GPS antennas were of the same manufacture. Little error was introduced because the variations canceled to first order.[55] Now that numerous GPS antennas are being used for precision geodetic research, phase center variations corrections are required.[46] If this is not done correctly, an error with an elevation angle dependence becomes apparent. This will be incorrectly modeled as part of the wet troposphere delay, which also exhibits elevation angle dependence. The end result is degraded vertical accuracy.

Ambiguity resolution is important for the most accurate and precise determinations of station coordinates using GPS. Ambiguity resolution converts a precise, yet biased, phase observable into an unbiased range observable of the same precision. Work published by Blewitt[3] and Dong and Bock[17] both find a factor of ~2.5 improvement in baseline precision of the eastern component due to ambiguity resolution for baselines from 50 to 500 km. They also suggest that ambiguity resolution improves agreement with VLBI. Network design is particularly important for successful ambiguity resolution. The confidence limit for resolving an ambiguity is dependent on the baseline length between the two receivers that formed the double difference. Ambiguities are resolved sequentially, meaning the solution covariance matrix is updated with the new information after each ambiguity is resolved. Being able to resolve the first few ambiguities may trigger successful resolution of the entire network. Thus, to ensure successful ambiguity resolution, an experiment coordinator might artificially introduce several short, ~10 km baselines. High-quality dual-frequency pseudorange is helpful for ambiguity resolution.

The final error source that requires attention is the GPS orbit. To first order, the orbit error dr maps onto baseline errors dx as

$$\frac{|\mathrm{d}x|}{x} \cong a \frac{|\mathrm{d}r|}{r} \tag{1}$$

where x is the baseline length, r is the altitude of the satellites, and a is a constant. The constant a has been found to have a value of approximately 0.2.[32] Apparently, this is because orbit errors are highly correlated, and this cancels some of the magnitude of their effects. Further reduction of the GPS orbit error depends on the ability of the user to define an accurate reference frame and on the sophistication of the models available.

The principles of orbit determination are simple and have been discussed in many textbooks. The true orbit of the GPS satellite differs from a pure Keplerian ellipse because of several perturbing forces, which include nonsphericity of the Earth's gravity field, attraction of the Sun and Moon, atmospheric drag, solar radiation pressure, and tides. The effects of atmospheric drag are negligible at the altitude of GPS satellites. The accelerations caused by the masses of the Sun and Moon are well known, and models can be used to describe the Earth's gravity field. Models for solar radiation pressure have been developed for both Block I and Block II satellites.[19] For arcs longer than ~12 h, solar radiation pressure bias parameters must be estimated. The limiting errors of GPS orbit determination are solar pressure and thermal radiation effects on the satellites themselves, particularly those satellites that are eclipsing.

III. Reference Frames

The fundamental aim of positional geodesy is to determine the location of a point on the Earth. The position of this point consists of three coordinates in a well-defined and accessible coordinate system. One common definition of a point on the Earth is the latitude, longitude, and altitude. Embedded in any definition of position is the concept of a reference frame. A Cartesian frame is useful for illustrative purposes. We seek to determine the location of a point in space; i.e., we seek to define a vector position $\boldsymbol{r}_e$. The coordinates of $\boldsymbol{r}_e$ must be referred to an origin. Additionally, we must define the orientation of the three-dimensional orthogonal axes. Finally, the scale, or vector length, must be defined. Thus, a total of seven parameters are required to define a reference frame: three terms for the origin, three terms for the orientation of the coordinate axes, and a scale. For a Cartesian terrestrial reference frame, the coordinate system origin is placed at the Earth's center of mass, or geocenter. The z axis by convention is aligned along the rotation axis. The x and y axes are orthogonal to the z axis, with the x axis traditionally defined at Greenwich.

If we want to measure changes in the position of a point on the Earth over time, we must ensure that what we measure is in fact the motion of the point fixed in the Earth's crust, and not motions associated with the reference frame. Thus, we seek to refer our geodetic measurements to an inertial frame. The motion of the Earth relative to the fixed stars, presumed to be an inertial reference frame, is well understood. Precession is the slow motion of the Earth's pole with respect to inertial space—with a period of approximately 26,000 years. Nutation refers to the oscillations of the pole over shorter periods. In addition, the Earth is rotating at a variable rate, and the pole of rotation moves. We refer to the variation in rotation as $UT1$ or U. The X and Y pole positions are defined by the rotation matrix p. The transformation between the inertial vector $\boldsymbol{r}_i$ and the Earth-fixed vector $\boldsymbol{r}_e$ can then be defined as a combination of these four individual transformations:

$$\boldsymbol{r}_i = PNUp\,\boldsymbol{r}_e \tag{2}$$

where N and P are nutation and precession, respectively. A more detailed discussion of these transformations can be found in Lambeck[27] or Vanicek and Krakiwsky.[52] Because all coordinates in these systems depend on the rotation of the Earth, one of the critical undertakings of geodesy is the study of variations of the Earth's rotation rate and motion of its axis of rotation.

Although only seven parameters are required to define the reference frame, in practice, more information is needed to realize this reference frame. A terrestrial reference frame is, by convention, defined by thousands of observations of station positions over years, such as has been determined by the International Earth Rotation Service in Paris using VLBI and SLR. The number of constraints that will be required to define the reference frame for the GPS analyst will be strongly dependent on the strength of the GPS constellation and the ground-tracking network.

For GPS geodesy, the reference frame is strongly linked to the issue of accurate orbit determination for the GPS satellites. As the GPS constellation and global tracking network have grown, different strategies have been developed. In its initial experimental Block I phase, only seven satellites were visible, over short periods of time. The GPS orbits were phased to favor tracking in the southwest United States. Figure 1 displays the satellite tracks visible in the late 1980s centered over Southern California. For regional (<500 km) experiments in Southern California, it was found that the orbits could be adequately determined by fixing the positions of just three receivers, or nine parameters. Thus, the positions of all other observing GPS receivers were estimated relative to the fixed sites. Because the fixed receivers determined the accuracy of the estimated positions, the receivers were often called a fiducial or "truth" network.[13] Figure 2 shows a typical fiducial network used in the 1980s for measuring crustal motions in a Southern California regional network. The receivers at Westford, Richmond, and Goldstone were operated by CIGNET (Cooperative International GPS Network). The coordinates of each site were determined by collocating the GPS receiver with VLBI or SLR monuments nearby. While the GPS constellation shown in

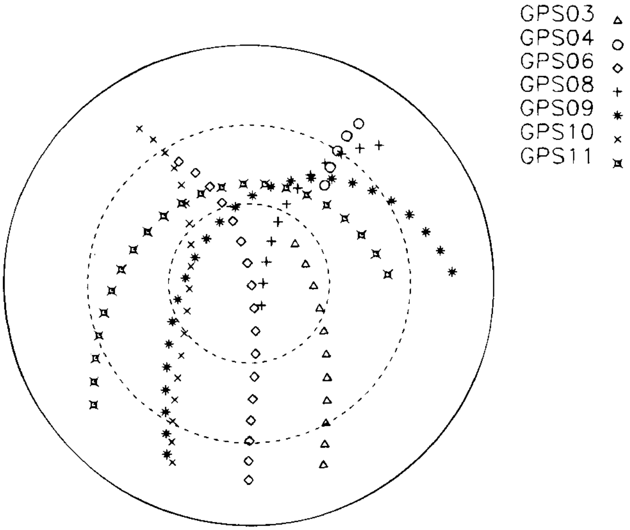


Fig. 1 Block I GPS sky tracks, centered over Southern California. These seven satellites were visible for a period of approximately 7 h. The dashed lines represent elevation angles of 30 and 60 deg, and the observations have been cut off at 15 deg.

Fig. 1 has a pronounced North–South orientation, the fiducial network has no such preference. Lichten et al.[34] and Larson et al.[30] both found that errors in the fiducial network, either through geometry or scale, could produce an appreciable systematic error in the network solution. If the coordinates of all available fiducial sites were known perfectly, of course, more than three sites should be fixed. Because of collocation errors, the coordinates of fiducial sites available to the GPS analyst in the late 1980s were only known with an accuracy of 3–4 cm. Because errors at fiducial sites directly affect the accuracy of the estimated receiver positions, it was desirable to fix the minimum number of sites required.

As GPS experiments became larger and more ambitious, ground tracking networks expanded. The 1988 Central and South American (CASA) experiment attempted to measure plate motions over distances of more than a 1000 km. A fiducial network centered over North America of approximately the same scale as the "regional" network would be insufficient. Thus, an international cooperative effort resulted in a global tracking network which extended to Europe and Australasia. This resulted in improvements in orbit accuracy and baseline precision.[26] Eventually CIGNET was expanded in Europe, Australia, New Zealand, and Japan through international cooperation. Other important sites in the southern hemisphere (Chile and South Africa) were added as part of the TOPEX/Poseidon precise orbit determination network. The global tracking network was expanded for the GPS experiment for the International Earth Rotation Service (IERS), also known as the GIG '91 experiment, as shown in Fig. 3. The GPS constellation consisted of 15 satellites at the time. The experiment lasted three weeks. Blewitt et al.[6] used data from GIG '91 to demonstrate a recent development in defining the GPS reference frame. Instead of fixing the coordinates of three or more sites,

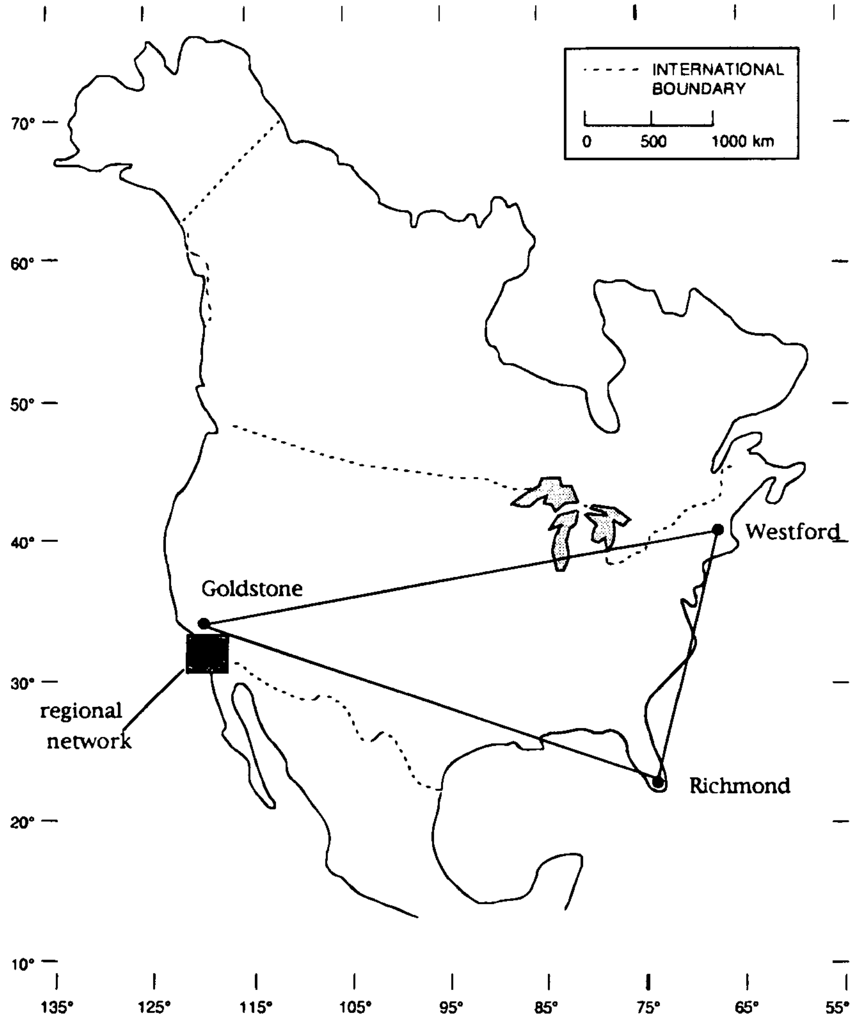


Fig. 2 Fiducial network for North American geodetic studies of the late 1980s. Cooperative International Network provided permanent GPS receivers at Westford, Goldstone, and Richmond. The Canadian Geodetic Survey provided data from Yellowknife.

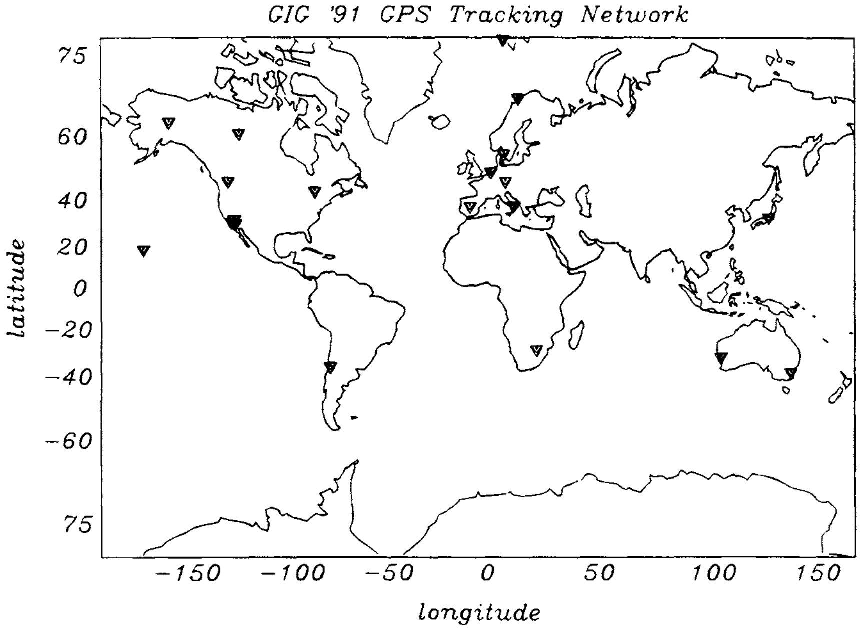


Fig. 3 Global tracking network for the GPS experiment for the International Earth Rotation Service and geodynamics.

they estimated all receiver coordinates. The scale and origin of the reference frame are implied by the satellite force model, propagation model, and the observables.[23] After estimation, Blewitt et al.[6] mapped the resulting station coordinates using a seven-parameter transformation into a frame defined by the ITRF (International Terrestrial Reference Frame),[10] which was itself derived from a joint VLBI-SLR analysis. They achieved accuracy levels commensurate with the errors in the ITRF. Because GIG'91 included 13 precisely determined globally distributed sites, Blewitt et al.[6] could use the information of all 13 sites, rather than only 3 or 4 sites.

The global GPS tracking network continues to expand on a monthly basis. Figure 4 shows some of the tracking stations that provide data to the IGS (International GPS Service for Geodynamics) at the present time. These data are available to all interested users through internet. The positions of these sites are also available from ITRF, derived from combined VLBI, SLR, and GPS estimates. *X* and *Y* pole positions and *UT*1 are distributed by IRIS (International Radio Interferometric Surveying) and IERS (International Earth Rotation Service). If desired, an analyst can retrieve GPS data from the IGS archives and estimate GPS orbits for the appropriate time period. Many organizations are also making precise GPS ephemerides available through the IGS. These orbits can be used to determine the relative positions of regional sites of interest. Because the GPS orbits are determined in the ITRF, the regional sites will also be defined in the correct reference frame, but without the need to estimate the satellite state.

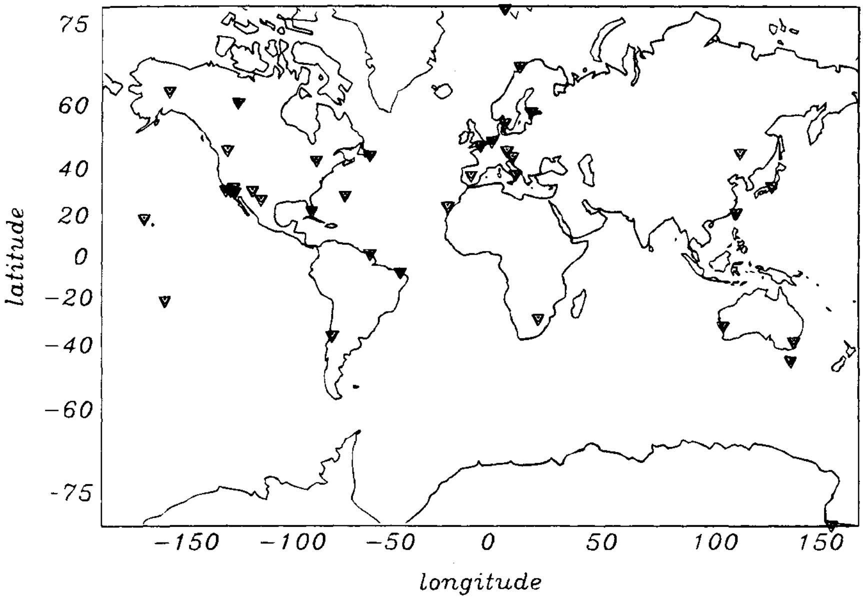


Fig. 4 GPS global tracking network, 1993.

Depending on the accuracy required, the analyst could also use a smaller fiducial network of three or four sites to determine the GPS orbits.

IV. Precision and Accuracy

Because its widescale use for precise geodesy began 10 years ago, GPS has quickly equaled the accuracy of mobile VLBI and SLR over regional and continental scales.[29,43] With the expansion of the constellation and improved modeling, global scale baselines also compare favorably with SLR and VLBI.[1,6]

The limits on baseline precision are determined by the geometry of the GPS constellation, strength of the tracking network, ability to model or correct error sources discussed in the previous section, and measurement noise. Refer again to Fig. 1, which displays the satellite constellation over Southern California in the late 1980s. The most precisely determined component was the North–South component, simply because more GPS sky tracks were North–South, with less variation in the East–West direction. Each site will have its skyplot characteristics, which are determined by the site latitude. Figure 5 shows the tracks for a site at the equator, with an extreme North–South preference, but without a hole in coverage between N30°W and N30°E, as appears over Southern California. Figure 6 shows the GPS sky tracks over the permanent GPS receiver at McMurdo station in Antarctica (77°S, 166°E). In this case, the satellite tracks are not preferentially aligned in either the East–West of North–South directions. The major limitation is that there are no satellite tracks whatsoever above an elevation angle of 60 deg.

Vertical precision is limited by skyview: we can look up but not down. The vertical component is also more sensitive to errors in the atmospheric path delay.

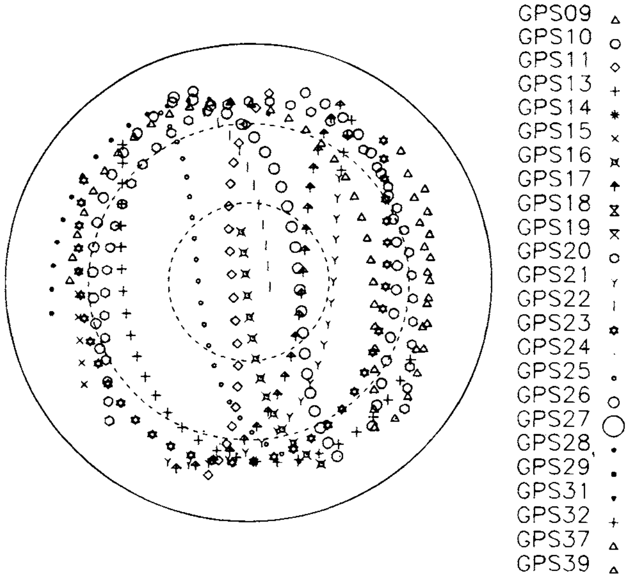


Fig. 5 GPS sky tracks centered at the equator over 12 h, August 1993.

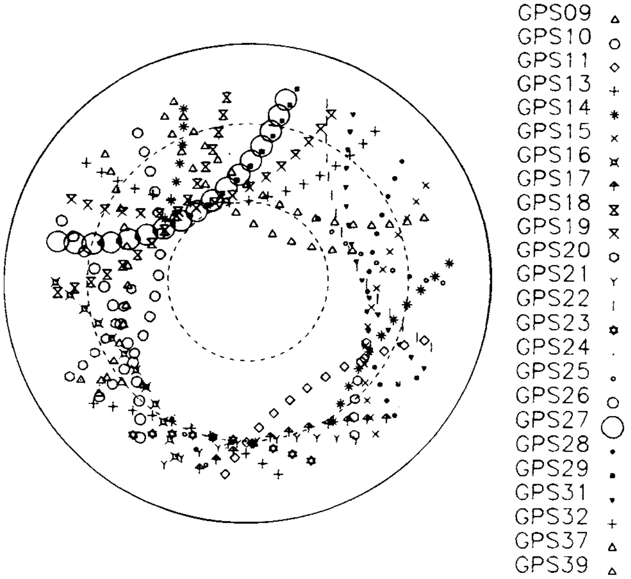


Fig. 6 GPS sky tracks centered over McMurdo Station (−77°S and 166°E), over 12 h, August 1993.

In order to differentiate between the atmospheric error and the vertical station position, a wide range of elevation angles should be observed. (A more detailed discussion of this issue can be found in Yunck.[56])

Precision is generally assessed by our ability to repeat a measurement, and thus it is often referred to as "repeatability." The repeatabilities are simply the weighted rms scatter about the weighted mean value. Because certain systematic errors do not manifest themselves at periods of a few weeks, we might expect that "short-term" repeatability, based on measurements over a few days, would be significantly better than "long-term" repeatability, based on measurements over many seasons and years. Because we are interested in system precision, unless qualified, the precision results we quote are long-term repeatabilities.

Repeatability for a regional network, where baselines spanned 50–350 km, in California was shown to be 2 mm + 6 × 10^{-9} L, for the north component, 2 mm + 13 × 10^{-9} L for the east component and 17 mm for the vertical, where L represents baseline length.[29] Davis et al.[14] studied baselines from 100 m to 225 km with comparable results. Over continental scales, Lichten and Border[32] found precisions of 0.6 and 2.6 cm for the North and East components, respectively, for a 1300-km baseline.

Although important, these early assessments of precision are being revised as the GPS constellation changes. The preliminary analyses all relied on the Block I constellation with four channel receivers and a maximum of 8 h of tracking. With a global network of receivers, a full constellation, and eight channel receivers, all visible satellites can be tracked simultaneously at a large number of sites. Under these new conditions, the greatest improvement in precision has been seen in the vertical component. With four channel receivers, the zenith troposphere delay parameter was highly correlated with the vertical component. This correlation is significantly reduced when the number of observed satellites increases from four to five or six.[5] Recent results from a permanent array of receivers spaced ~200 km apart indicate horizontal precisions of 4 mm and vertical precisions of 10–15 mm.[35] Over longer baselines, Heflin et al.[23] report baseline length repeatabilities of 2 mm + 4 × 10^{-9} L, up to 12,000 km. Anderson et al.[1] find similar precision over global scales.

Accuracy of GPS is determined by comparison with a truth standard. Most accuracy assessments for GPS have consisted of comparisons of baseline components with VLBI. Since the VLBI reference frame has been adopted for the fiducial coordinates used to define the GPS reference frame, these comparisons are self-consistent. If the VLBI and GPS observations had been made at the same ground monument, the comparison would be fairly simple to carry out. Unfortunately, the uncertainty in survey ties between VLBI and GPS monuments often corrupt the comparison. Although blunders, say, 10 cm, are fairly easy to determine, small systematic errors of several centimeters' magnitude can be incorrectly attributed to the new measurement system. Lichten and Bertiger[33] compared continental scale baselines, finding an agreement of 1.5 parts in 10^8 for 2000-km baselines. Suggesting that some of the VLBI–GPS comparisons in California were contaminated with survey errors, Larson and Agnew[29] instead compared four baseline rates and found agreement to better than one standard deviation. Davis et al.[14] compared electronic distance measurements to GPS and found no significant differences between the two systems.

With the permanent, continuously observing, global tracking network, there have been further improvements in precision and accuracy of GPS geodesy. Blewitt et al.[8] report absolute station coordinate accuracy of 15 mm, with weekly repeatabilities of better than 5 mm in latitude and longitude, and 10 mm in height. Furthermore, the independently determined GPS scale agrees with both VLBI and SLR to within 1 part in 10^9, well within the uncertainties of all three systems.[6] These Jet Propulsion Laboratory (JPL) results are summarized in Fig. 7.

V. Results

A. Crustal Deformation

Much of the interest in the geodetic accuracy of GPS is driven by scientists investigating deformation of the Earth's surface. Measurements of the Earth's surface provide direct tests of geophysical models used to describe the dynamics of the Earth. One important hypothesis tested by geodetic measurements is the theory of plate tectonics. Although sea-floor rocks that recorded magnetic reversals occurring over time scales of millions of years provided the first strong evidence of this phenomenon, contemporary measurements by space geodetic systems have shown that those long-term rates are statistically consistent with rates determined over a decade.[11,50] Other fundamental questions that should benefit from geodetic surveys include identification of earthquake hazard zones, mountain-building processes, and estimates of mantle viscosity via accurate measurements of postglacial isostatic rebound.

Nearly all crustal deformation surveys to date have concentrated on determining the horizontal deformation rates of the Earth's crust. This has been particularly successful across transform boundaries, such as the San Andreas Fault (SAF) in

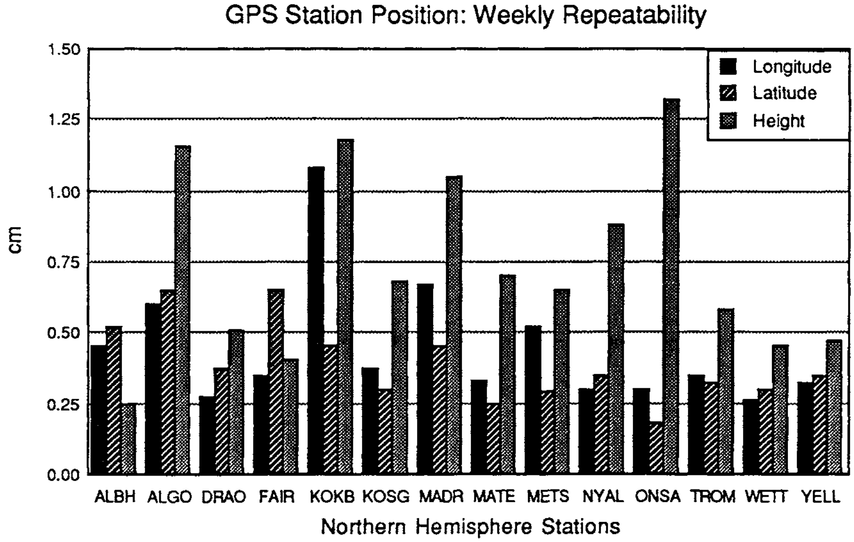


Fig. 7a Weekly repeatability of 14 northern hemisphere sites over 13 weeks during the summer of 1992. The typical daily repeatability is 3–5 mm in the horizontal and 5–10 mm in the vertical.

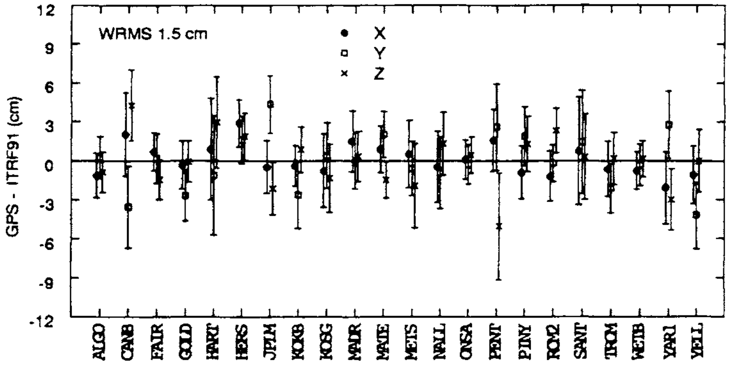


Fig. 7b Comparisons with independent very long baseline interferometry and satellite laser ranging measurements indicate an absolute 1-sigma accuracy of about 1.5 cm.

California, where no vertical deformation is expected. Prescott et al.[44] discussed GPS derived baselines that crossed the SAF in what became the rupture zone of the 1989 Loma Prieta earthquake. The NW–SE relative station velocities, characteristic of the SAF in California, were apparent in the three years of measurements conducted by the U.S. Geological Survey. Deformation rates across the SAF obtained with GPS were consistent with measurements made by both electronic distance measurements (trilateration)[37] and VLBI.[11]

GPS has also been used to address other important tectonic questions. Following the controversy regarding the magnitude of the Pacific-North American plate rate, Dixon et al.[16] began making GPS measurements across the Gulf of California. Over three and one-half years, they estimate a plate rate of 47 ± 7 mm/yr at a direction of −57 ± 6 deg. Their contemporary measurements agree better with the NUVEL1 global plate prediction[15] than with the previous standard plate motion model.[42] Although nearly 75% of the Pacific–North American relative plate motion is taken up on the SAF, it is not understood where the remaining portions of the deformation are distributed. Characterizing the deformation to the west of the San Andreas Fault has been the focus of a recent collaborative effort.[18] Their results are summarized in Fig. 8. Larson and Webb[31] found deformation rates on the order of 5 mm/yr across the eastern Santa Barbara Channel, which is consistent with seismic evidence of deformation.

In Japan, Shimada and Bock[48] were able to monitor the convergence of the Eurasian, Pacific, North American, and Philippine Sea plates, yielding westward motion of 28 mm/yr and significant vertical uplift. Freymueller et al.[21] describe the interactions of the Nazca, Cocos, Caribbean, and South American plates. As shown in Fig. 9, the geodetic results are in good agreement with NUVEL-1, although there are significant differences. Meertens and Smith[41] report on deformation associated with the Yellowstone Caldera.

Many geophysicists are also interested in the potential of GPS for measuring accurate vertical components. To date, there have been no significant estimates of secular vertical rates, such as postglacial isostatic rebound, with GPS although the installation of tracking stations by the Canadian Geodetic Survey at Penticton,

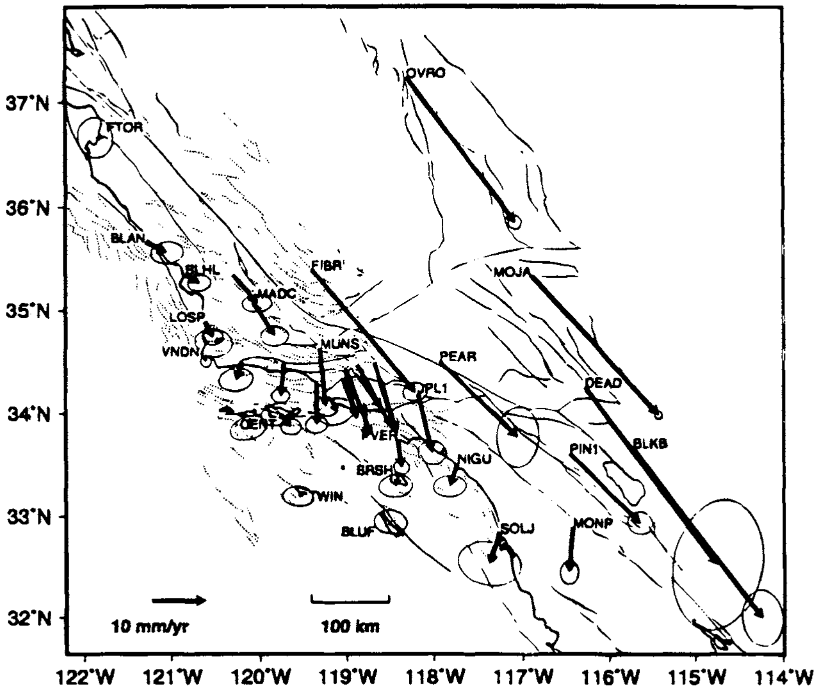


Fig. 8 **Observed velocity of stations relative to the Pacific plate estimated from the combined global positioning system and very long baseline interferometry data set. The ellipses denote the region of 95% confidence, after scaling the formal uncertainties by a factor of 2. For clarity, the ellipses are not shown for the sites in the Ventura Basin. Reprinted from Feigl et al.[18] with the permission of the American Geophysical Union.**

Algonquin, St. Johns, and Yellowknife should provide important constraints in the coming years.

Abrupt motions, such as those associated with earthquakes and volcanoes, have also been measured with GPS. Larsen et al.[28] analyzed data collected in the Imperial Valley of Southern California before and after the 1987 Superstition Hills earthquake, yielding a detailed map of the surface displacements. They were able to model surface displacements into right lateral slip of 130 ± 8 mm and 30 ± 10 mm of left-lateral slip. The 1991 Costa Rican earthquake (surface wave magnitude of 7.6) caused displacements of up to 2.4 m.[38] Analysis of GPS data was useful in determining a new dislocation model for the earthquake. Displacements associated with the 1992 Landers earthquake were the subject of work by both the JPL[7] and a group of university researchers.[9] These displacements are now available to the geodetic community to improve geodetic control in the Southern California region. Magma chamber deflation associated with the 1991 Hekkla volcanic eruption was discerned by repeated GPS measurements discussed by Sigmundssen.[49]

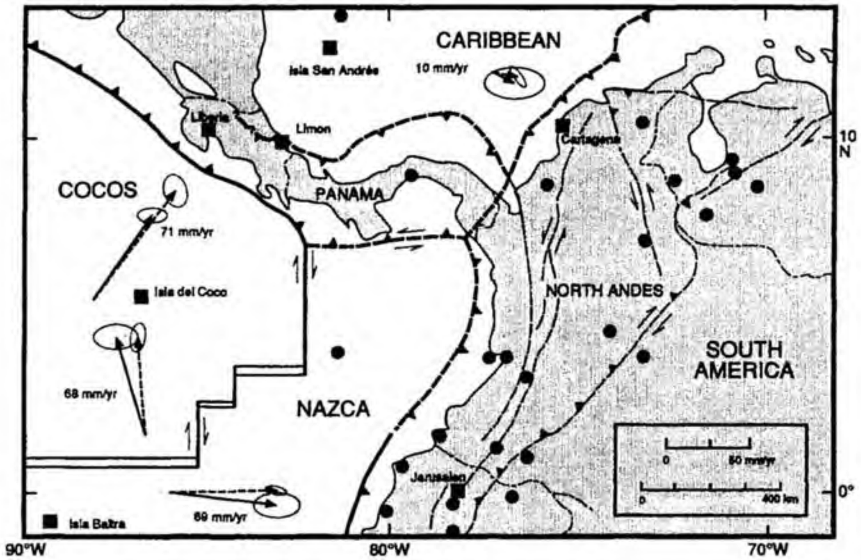


Fig. 9 Observed GPS (solid lines) and model NUVEL-1[15] (dashed lines) baseline rates of change, with their 95% confidence ellipses. Baseline rates of change measure the relative motions of two GPS sites. The Caribbean–North Andes plate motion is based on only two epochs of GPS data (1988 and 1991), and the uncertainty given is conservative. GPS sites used in this study are indicated by squares; other GPS sites are indicated by circles. Reprinted from Freymueller et al.[20] with the permission of the American Geophysical Union.

B. Earth Orientation

The most recent precise geodetic contribution by GPS has come about with the advent of global tracking networks established in the early 1990s. Global tracking networks and a larger GPS constellation now yield sufficient sensitivity to allow estimation of the geocenter and polar motion. One goal of the GIG '91 experiment discussed in Sec. III was to investigate the potential of GPS to determine Earth rotation parameters. A 21-station subset of the GIG '91 network was analyzed by independent groups. The GPS constellation consisted of 15 satellites at the time. Groups at MIT and JPL were able to estimate daily X and Y pole positions to an accuracy of better than 1 milliarcsec.[24,36] Variations in $UT1$ were estimated by Lichten et al.[34] with an accuracy of a few hundredths of a millisecond.

The geocenter was the subject of a companion study by JPL.[53] They found an offset with respect to ITRF of −8.3 ± 2.7, 13.4 ± 2.4, and −7.7 ± 13.7 cm in X, Y, and Z, respectively. With an expanded constellation and global tracking network (particularly in the southern hemisphere) in 1992, Malla et al.[40] reported much improved X and Y components, 0.0 ± 1.4, 1.5 ± 1.3 cm, but the Z estimate was still weak, −8.2 ± 3.0 cm. With the addition of TOPEX GPS data, Malla[39] was able to improve the Z component agreement with ITRF to 0.1 ± 1.5 cm.

VI. Conclusions

Geodetic science has greatly benefited from the GPS. With the appropriate analysis strategy and reference frame, the geodesist can now determine absolute geocentric positions with an accuracy of better than 2 cm, with even greater accuracy for relative positions. Permanent GPS tracking sites can now be used to maintain the terrestrial reference frame, including subdaily resolution of Earth orientation parameters. Within the global network, geophysicists will be able to measure plate motions and test assumptions of plate rigidity. Regional "mobile" GPS experiments will be used to study complicated seismic zones, such as the Himalayan collision zone, uplift in Chile, and subduction in Japan. Within five years there should be significant estimates of postglacial rebound from monitoring programs begun in Canada and Fennoscandia. Finally, GPS will play an important role in monitoring sea level by tieing tide gauges into the global terrestrial reference frame.

Acknowledgments

This manuscript was begun while I was a visiting scientist at the Branch of Earthquake Geology and Geophysics of the U.S. Geological Survey and completed at the Laboratory for Terrestrial Physics of the NASA Goddard Space Flight Center. The work is supported by NASA NAG 5-1908 and NSF EAR-9209385. I am grateful to Robert B. Miller, Geoff Blewitt, Jim Davis, Jeff Freymueller, Steve Nerem, George Rosborough, and Mark Tapley for helpful discussions. I thank Tom Yunck and Steve Lichten for reviewing the manuscript and helping me improve it.

References

[1]Anderson, P. H., Hauge, S., and Kristiansen, O., "GPS Relative Positioning at a Precise Level of One Part per Billion," *Bulletin Geodesique,* Vol. 67, 1993, pp. 91–106.

[2]Beutler, G., Bauersima, I., Gurtner, W., Rothacher, M., and Schildknecht, T., "Evaluation of the 1984 Alaska Global Positioning System Campaign with the Bernese GPS Software," *Journal of Geophysical Research,* Vol. 92, 1987, pp. 1295–1303.

[3]Blewitt, G., "Carrier Phase Ambiguity Resolution for the Global Positioning System Applied to Geodetic Baselines up to 2000 km," *Journal of Geophysical Research,* Vol. 94, 1989, pp. 10,187–10,283.

[4]Blewitt, G., "An Automatic Editing Algorithm for GPS Data," *Geophysical Research Letters,* Vol. 17, 1990, pp. 199–202.

[5]Blewitt, G., "Advances in Global Positioning System Technology for Geodynamics Investigators, 1978–1992," *Contributions of Space Geodesy: Technology,* edited by D. Smith and D. Turcotte, Vol. 25, Geophysical Monograph Series, American Geophysical Union, Washington, DC, 1993, pp. 195–213.

[6]Blewitt, G., Heflin, M. B., Webb, F. H., Lindqwister, U. J., and Malla, R. P., "Global Coordinates with Centimeter Accuracy in the International Terrestrial Reference Frame using the Global Positioning System," *Geophysical Research Letters,* Vol. 19, 1992, pp. 853–856.

[7]Blewitt, G., Heflin, M. B., Hurst, K. J., Jefferson, D. C., Webb, F. H., and Zumberge, J. F., "Absolute Far-Field Displacements from the 28 June 1992 Landers Earthquake Sequence," *Nature,* Vol. 361, 1993, pp. 340–342.

[8]Blewitt, G., Heflin, M. B., Hurst, K., Jefferson, D., Vigue, Y., Webb, F., and Zumberge, J., "Viewing the Earth as a Rotating, Deforming Polyhedron, Using the Global Positioning System," *EOS,* Vol. 74, No. 16, 1993, p. 48.

[9]Bock, Y., et al., "Detection of Crustal Deformation from the Landers Earthquake Sequence Using Continuous Geodetic Measurements," *Nature,* Vol. 361, 1993, pp. 338–340.

[10]Boucher, C., and Altamimi, Z., "ITRF 89 and other Realizations of the IERS Terrestrial Reference System for 1989," *IERS Technical Note 6,* Observatoire de Paris, 1991.

[11]Clark, T. A., Gordon, D., Himwich, W. E., Ma, C., Mallana, A., and Ryan, J. W., "Determination of Relative Site Motions in the Western United States Using Mark III Very Long Baseline Interferometry," *Journal of Geophysical Research,* Vol. 92, 1987, pp. 12,741–12,750.

[12]Cohen, C., and Parkinson, B., "Mitigating Multipath Error for GPS-Based Attitude Determination," *Keystone Guidance and Contral Conference, Proceedings of the American Astronomical Society,* (Keystone, CO), Feb. 1991.

[13]Davidson, J. M., Thornton, C. L., Vegos, C. J., Young, L. E., and Yunck, T. P., "The March 1985 Demonstration of the Fiducial Network for GPS Geodesy: A Preliminary Report," *Proceedings of the First International Symposium on Precise Positioning with GPS,* edited by C. Goad, NOAA, Rockville, MD, 1985, pp. 603–612.

[14]Davis, J. L., Prescott, W. H., Svarc, J. L., and Wendt, K. J., "Assessment of Global Positioning System Measurements for Studies of Crustal Deformation," *Journal of Geophysical Research,* Vol. 94, 1989, pp. 13,635–13,650.

[15]DeMets, C., Gordon, R., Argus, D., and Stein, S., "Current plate motions," *Geophysical Journal International,* Vol. 101, 1990, pp. 425–478.

[16]Dixon, T. H., Gonzalez, G., Lichten, S., Tralli, D., Ness, G., and Dauphin, J., "Preliminary Determination of Pacific-North America Relative Motion in the Southern Gulf of California using the Global Positioning System," *Geophysical Research Letters,* Vol. 18, 1991, pp. 861–864.

[17]Dong, D., and Bock, Y., "GPS Network Analysis with Phase Ambiguity Resolution Applied to Crustal Deformation Studies in California," *Journal of Geophysical Research,* Vol. 94, 1989, pp. 3949–3966.

[18]Feigl, K., et al., "Space Geodetic Measurement of Crustal Deformation in Central and Southern California," *Journal of Geophysical Research,* Vol. 98, 1993, pp. 21, 677–621, 712.

[19]Fliegel, H. F., Gallini, T. E., and Swift, E. R., "Global Positioning System Radiation Force Model for Geodetic Applications," *Journal of Geophysical Research,* Vol. 97, 1991, pp. 559–568.

[20]Freymueller, J., Kellogg, J., and Vega, V., "Plate Motions in the North Andean Region," *Journal of Geophysical Research,* Vol. 98, 1993, pp. 21, 853–821, 863.

[21]Freymueller, J., "Phasedit—A GPS Cycle Slip Detector," *IOM,* Stanford University, Stanford, CA, 1993.

[22]Genrich, J. F., and Bock, Y., "Rapid Resolution of Crustal Motion with Short-Range GPS," *Journal of Geophysical Research,* Vol. 97, 1992, pp. 3261–3270.

[23]Heflin, M. B., et al., "Global Geodesy Using GPS without Fiducial Sites," *Geophysical Research Letters,* Vol. 19, 1992, pp. 131–134.

[24]Herring, T. A., Dong, D., and King, R. W., "Sub-Milliarcsecond Determination of Pole Position Using Global Positioning System Data," *Geophysical Research Letters,* Vol. 18, 1991, pp. 1893–1986.

[25]King, R. W., Masters, E. G., Rizos, C., Stolz, A., and Collins, J., *Surveying with GPS,* School of Surveying Monograph 9, University of New South Wales, Kensington, Australia, 1985.

[26]Kornreich-Wolf, S., Dixon, T. H., and Freymueller, J. T., "The Effects of Tracking Network Configuration on GPS Baseline Estimates for the CASA UNO Experiment," *Geophysical Research Letters,* Vol. 17, 1990, pp. 647–650.

[27]Lambeck, K., *Geophysical Geodesy, the Slow Deformations of the Earth,* Clarendon, Oxford, 1988.

[28]Larsen, S., Reilinger, R., Neugebauer, H., and Strange, W., "Superstition Hills Earthquake," *Journal of Geophysical Research,* Vol. 97, 1992, pp. 4885–4902.

[29]Larson, K. M., and Agnew, D., "Application of the Global Positioning System to Crustal Deformation Measurement: 1. Precision and Accuracy," *Journal of Geophysical Research,* Vol. 96, 1991, pp. 16,547–16,565.

[30]Larson, K. M., Webb, F. H., and Agnew, D., "Application of the Global Positioning System to Crustal Deformation Measurement: 2. The Influence of Errors in Orbit Determination Networks," *Journal of Geophysical Research,* Vol. 96, 1991, pp. 16,567–16,584.

[31]Larson, K. M., and Webb, F. H., "Active Deformation in the Santa Barbara Channel Inferred from GPS Measurements," *Geophysical Research Letters,* Vol. 19, 1992, pp. 1491–1494.

[32]Lichten, S. M., and Border, J. S., "Strategies for High-Precision Global Positioning System Orbit Determination," *Journal of Geophysical Research,* Vol. 92, 1987, pp. 12,751–12,762.

[33]Lichten, S. M., and Bertiger, W. I., "Demonstration of Sub-Meter GPS Orbit Determination and 1.5 Parts in 10^8 Three-Dimensional Baseline Accuracy," *Bulletin Geodesique,* Vol. 63, 1989, pp. 167–189.

[34]Lichten, S. M., Bertiger, W. I., and Lindqwister, U. J., "The Effect of Fiducial Network Strategy on High-Accuracy GPS Orbit Determination and Baseline Determination," *Proceedings of the 5th International Symposium on Satellite Positioning,* (Las Cruces, NM), 1989.

[35]Lindqwister, U. J., Zumberge, J. F., Webb, F. H., and Blewitt, G., "Few Millimeter Precision for Baselines in the California Permanent GPS Geodetic Array," *Geophysical Research Letters,* Vol. 18, 1991, pp. 1,135–1,138.

[36]Lindqwister, U. J., Freedman, A. P., and Blewitt, G., "Daily Estimates of the Earth's Pole Position with the Global Positioning System," *Geophysical Research Letters,* Vol. 19, 1992, pp. 845–848.

[37]Lisowski, M., Savage, J. C., and Prescott, W. H., The Velocity Field Along the San Andreas Fault in Central and Southern California, *Journal of Geophysical Research,* Vol. 96, 1991, pp. 8369–8389.

[38]Lundgren, P., Kornreich Wolf, S., Protti, M., and Hurst, K. J., "GPS measurements of crustal deformation associated with the 22 April 1991, Valle de la Estrella, Costa Rica earthquake," *Geophysical Research Letters,* Vol. 20, 1993, 407–410.

[39]Malla, R. P., "Breaking the ΔZ Barrier in Geocenter Estimation," Jet Propulsion Laboratory, JPL Interoffice Memorandum 335.8-93-018, Aug. 18, 1993.

[40]Malla, R. P., Wu, S. C., and Lichten, S. M., "Use of Global Positioning System Measurements to Determine Geocentric Coordinates and Variations in Earth Orientation," NASA TDA Progress Rept. 42-114, 1993.

[41]Meertens, C. M., and Smith, R. B., "Crustal Deformation of the Yellowstone Caldera from first GPS Measurements: 1987–1989," *Geophysical Research Letters,* Vol. 18, 1991, pp. 1763–1766.

[42]Minster, J. B., and Jordan, T. H., "Present-day plate motions," *Journal of Geophysical Research,* Vol. 83, 1978, pp. 5331–5354.

[43]Murray, M. H., "Global Positioning System Measurement of Crustal Deformation in Central California," Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1991.

[44]Prescott, W. H., Davis, J. L., and Svarc, J. L., "Global Positioning System Measurements for Crustal Deformation: Precision and Accuracy," *Science,* 1989, pp. 1337–1340.

[45]Rocken, C., Johnson, J. M., Neilan, R. E., Cerezo, M., Jordan, J. R., Falls, M. J., Nelson, L. D., Ware, R. H., and Hayes, M., "The measurement of atmospheric water vapor, radiometer comparison and spatial variations," *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 29, 1991, pp. 3–8.

[46]Schupler, B. R., and Clark, T. A., "How Different Antennas Affect the GPS Observable," *GPS World,* Vol. 2, No. 10, 1991, pp. 32–36.

[47]Schutz., B. E., Tapley, B. D., Abusali, P. A. M., Rim, H. J., "Dynamic Orbit Determination Using GPS Measurements from TOPEX/POSEIDON," *Geophysical Research Letters,* (submitted for publication).

[48]Shimada, S., and Bock, Y., "Crustal Deformation Measurements in Central Japan Determined by a Global Positioning System Fixed-Point Network," *Journal of Geophysical Research,* Vol. 97, No. B9, 1992, pp. 12,437–12,456.

[49]Sigmundsson, F., Einarsson, P., and Bilham, R., "Magma chamber deflation recorded by the Global Positioning System: the Hekkla 1991 eruption," *Geophysical Research Letters,* Vol. 14, 1992, pp. 1483–1486.

[50]Smith, D. E., et al., "Tectonic Motion and Deformation from Satellite Laser Ranger to LAGEOS," *Journal of Geophysical Research,* Vol. 95, 1990, pp. 22,013–22,042.

[51]Tralli, D. M., Dixon, T. H., and Stephens, S. A., "The Effect of Wet Tropospheric Path Delays on Estimation of Geodetic Baselines in the Gulf of California Using the Global Positioning System," *Journal of Geophysical Research,* Vol. 93, 1988, pp. 6545–6557.

[52]Vanicek, P., and Krakiwsky, E. J., *Geodesy: The Concepts,* North Holland Publishing Company, Amsterdam, 1986.

[53]Vigue, Y., Lichten, S. M., Blewitt, G., Heflin, M. B., and Malla, R. P., Precise determination of the Earth's center of mass using measurements from the Global Positioning system, *Geophysical Research Letters,* Vol. 19, 1992, pp. 1487–1490.

[54]Ware, R., Rocken, C., Solheim, F., Van Hove, T., Alber C., and Johnson, J., "Pointed Water Vapor Radiometer Corrections for Accurate Global Positioning Surveying," *Geophysical Research Letters,* Vol. 20, 1993, pp. 2635–2638.

[55]Wu, J., Wu, S., Hajj, G., Bertiger, W., and Lichten, S. "Effects of Antenna Orientation on GPS Carrier Phase," *Astrodynamics 1991,* edited by B. Kaufman et al., Vol. 76, *Adv. in Astron. Sci.,* 1992, pp. 1647–1660.

[56]Yunck, T. P., "Coping with the Atmosphere and Ionosphere in Precise Satellite and Ground Positioning," *Environmental Effects on Spacecraft Trajectories and Positioning,* edited by A. Vallance-Jones, Geophysical Monograph 73, IUGG Vol. 13, 1993, pp. 1–16.

Chapter 21

Orbit Determination

Thomas P. Yunck*
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109

I. Introduction

AN Earth satellite collecting GPS data with an onboard receiver can compute its state (position and velocity) in a diversity of ways, the choice depending in part on the type of orbit and mission requirements. Tracking and navigation requirements can include real-time state knowledge and active control during launch and orbit insertion[1] and during re-entry and landing; real-time relative navigation between vehicles during rendezvous[2,3]; autonomous stationkeeping and near-real-time orbit knowledge for operations and orbit maintenance[4]; rapid postmaneuver orbit recovery[5]; and after-the-fact precise orbit determination for scientific analysis.[6,7] Orbit accuracy requirements can range from hundreds of meters or more for routine operations to a few centimeters for precise remote sensing. Among existing tracking systems, only GPS can meet the most stringent of these needs for the most dynamically unpredictable vehicles. An overview of GPS space applications is given in Ref. 8.

The GPS signal beamwidths extend roughly 3000 km beyond the Earth's limb, enabling an Earth orbiter below that altitude to receive continuous three-dimensional coverage. This chapter focuses on orbit estimation for satellites in low circular orbits, below a few thousand kilometers, with emphasis on the high accuracy that GPS so ably provides. Real-time techniques fall under what we call *direct* GPS orbit determination, in which only the GPS data collected by the orbiter are used in the solution. For precise after-the-fact solutions, we turn to a global form of *differential* GPS in which data collected at multiple ground sites are combined with the onboard data to reduce the major errors. We also examine briefly the adaptation of GPS tracking techniques to satellites in highly elliptical and geosynchronous orbits.

*Deputy Manager, Tracking Systems and Applications Section.

The potential of GPS to provide accurate and autonomous satellite orbit determination was noted early in its development.[9] Early studies of direct GPS-based tracking include those in Ref. 10, which surveyed applications from near Earth to beyond geosynchronous altitudes; Ref. 11, which examined GPS tracking of the Space Shuttle; Ref. 12, which focused on autonomous near Earth navigation; Ref. 13, which described NASA's first planned GPS orbital application to Landsat-4; Ref. 14, which compared the potential of GPS and NASA's Tracking and Data Relay Satellite System (TDRSS) for onboard navigation; Ref. 15, which discussed flight receiver requirements and expected onboard orbit accuracies from near Earth to geosynchronous altitude; and Ref. 16, which explored geosynchronous applications. The first reported results from direct GPS tracking were those of the Landsat-4 experiment,[17,18] which achieved approximately 20 m accuracy during the relatively brief periods of good GPS visibility at that time.

Among the first descriptions of precise orbit determination at the level of several decimeters or better by differential GPS techniques are those in Ref. 19, which proposed a subdecimeter carrier phase-based technique for the TOPEX (later TOPEX/Poseidon) ocean altimetry mission; Ref. 20, which examined differential tracking of a low-altitude orbiter; Ref. 21, which proposed differential techniques for high-altitude satellites; and Ref. 22, which surveyed a variety of differential GPS applications. Since then, several important refinements have been introduced which better exploit the unique signals and the unprecedented observing strength GPS provides.

II. Principles of Orbit Determination

Point positioning with GPS is as accurate in low orbit as on the ground: typically 50–100 m for the GPS standard positioning service (SPS) user (under nominal levels of selective availability) and 10–20 m for the precise positioning service (PPS) user. Corresponding velocity solutions from carrier phase rate may reach 0.5 m/s (SPS) and better than 0.1 m/s (PPS). Although those levels are adequate for many purposes, instantaneous solutions have their limitations. They may be impossible during periods of restricted visibility, for example, and their accuracy may be inadequate for orbit prediction or for some real-time needs. Some scientific instruments require real-time position knowledge of meters to tens of meters for accurate pointing, while after-the-fact requirements can be far more stringent. To reduce the instantaneous position and velocity error, the traditional tools of dynamic orbit estimation can be brought to bear.

A. Dynamic Orbit Determination

Classical dynamic orbit determination exploits orbital mechanics—the physics underlying orbital motion—and filtering theory to yield a stable and accurate orbit solution from generally sparse and noisy measurements. This approach has, in fact, been necessary with conventional tracking systems, which, unlike GPS, seldom if ever provide sufficient information at one time for a geometric solution, and can provide no data at all over much of the globe. (An exception is the use of range and angle data from a single site to determine the instantaneous position of geostationary satellites, although the accuracy of that technique is far worse than that of GPS.) An orbit model must, therefore, be introduced to supply the missing information. In dynamic orbit determination, the orbit model is derived from models of the forces acting on the satellite and the laws of motion.

The process begins with a set of tracking measurements (range or Doppler, for example) along with mathematical models of the forces acting on the satellite and of the satellite physical properties. The major forces include gravity, aerodynamic drag and lift, solar radiation pressure, and active thrusting. Lesser contributions may come from outgassing, satellite thermal radiation, sunlight reflected from the Earth, and electromagnetic effects. The force and satellite models are used to compute a model of satellite acceleration over time, from which, by double integration, a nominal trajectory is formed. In principle, all that is then needed to produce the orbit solution is to determine the two vector constants of integration—position and velocity at one time point—also known as the epoch state. That is done through an estimation procedure that finds the epoch state for which the resulting model trajectory best fits the tracking data according to some optimality criterion, usually minimizing the mean square fitting error. To improve the fit, we can simultaneously estimate various force parameters, such as drag, solar radiation, and gravity coefficients; geometrical parameters, such as tracking station positions and Earth rotation; or empirical parameters, such as nonspecific once- and twice-per-orbit accelerations. The resulting solution is still a trajectory derived from force models, and its accuracy depends on how faithfully those models, fixed or adjusted, describe the real forces acting on the satellite.

More formally, to construct a nominal or a priori satellite trajectory we begin with Newton's second law of motion

$$f = m\,a = m\,\ddot{\boldsymbol{r}} \tag{1}$$

or

$$\ddot{\boldsymbol{r}} = f/m \tag{2}$$

where $\boldsymbol{r}$ is the satellite position vector. This fundamental equation of mechanics provides the dynamical constraint governing the orbit solution. The true acceleration $\ddot{\boldsymbol{r}}$ at any instant depends on the satellite position and velocity at that instant, and on many other parameters that characterize the forces at work. In the orbit solution, those parameters may take the form of spherical harmonic gravity coefficients, drag and lift coefficients, solar flux and reflectivity, a geomagnetic index, and so on. Let $(\boldsymbol{r}_o, \dot{\boldsymbol{r}}_o)$ be the true satellite epoch state to be estimated. We first select a nominal epoch state $(\boldsymbol{r}_{on}, \dot{\boldsymbol{r}}_{on})$, perhaps from an instantaneous GPS state solution, and construct an acceleration model $\ddot{\boldsymbol{r}}_n\,(t)$ from the force and satellite models. The nominal trajectory $\boldsymbol{r}_n\,(t)$ is then generated by double integration of the acceleration model,

$$\boldsymbol{r}_n(t) = \iint \ddot{\boldsymbol{r}}_n(t)\,\mathrm{d}t + \dot{\boldsymbol{r}}_{on}\,t + \boldsymbol{r}_{on} \tag{3}$$

The least squares solution procedure will then estimate corrections to the nominal epoch state (and possibly to selected model parameters) that bring the model trajectory into better agreement with the tracking data. If only the six-element epoch state and a few other parameters are adjusted, as is commonly the case, then in principle only a relatively few measurements around the orbit are needed to yield a well-determined solution, and a sparse tracking network will suffice. This is the great power and appeal of dynamic orbit determination. Since the first days of space exploration this technique has made practical the accurate tracking of Earth satellites and deep space probes.

Observe, however, that the resulting orbit solution depends intimately on the (possibly adjusted) acceleration model $\ddot{r}_n(t)$. Where high-accuracy orbits are required, high-accuracy models must be found. This can be enormously costly and may be a practical impossibility for low-altitude or maneuvering vehicles. In the mid-1980s it was recognized that the continuous three-dimensional coverage given by GPS offers an escape from this dynamical bind. Before describing these orbit determination techniques, we first review some principles of optimal estimation theory.

B. Batch Least Squares Solution

An enduring technique for estimating celestial orbits is the method of least squares, first employed by Gauss in 1795. Let z be a vector of observations $(z_1, \ldots, z_n)^T$ made over an interval of time, or tracking arc. The objective is to find that trajectory, among all possible trajectories satisfying the dynamical constraint [Eq. (2)], which minimizes the mean square difference between the actual observations z_i and theoretical observations $\tilde{z}_i$ derived from the solution trajectory. That is, we want to find the trajectory $\boldsymbol{r}(t)$ that minimizes the functional

$$J = \sum_{i=1}^{n} \{z_i - \tilde{z}_i\,[\boldsymbol{r}(t)]\}^2 \tag{4}$$

As this is a nonlinear problem, we reformulate it as one of computing a linear correction to the nominal trajectory $\boldsymbol{r}_n(t)$ given by Eq. (3). First, we compute theoretical observations $\tilde{z}_i$ from the nominal trajectory, then form the differences $\delta z_i = z_i - \tilde{z}_i$. These *prefit residuals* become the observations to be used in a linear adjustment of the nominal trajectory. (Strictly speaking, this is still not a linear problem; but if the nominal trajectory is sufficiently close to the true trajectory, it will be in the "linear regime," where a linear correction is adequate, if not perfect. If greater accuracy is needed, a linear correction to the new solution can be computed, and so on for multiple iterations, until the solution converges.) The familiar linear equation can be written as follows

$$\delta z = A\,\boldsymbol{x} + \boldsymbol{n} \tag{5}$$

where $\boldsymbol{x}$ is the vector of parameters to be estimated, $\boldsymbol{n}$ is the vector of random measurement noise on the observations δz, and A is a matrix of partial derivatives of the observations with respect to the elements of $\boldsymbol{x}$. Here $\boldsymbol{x}$ includes, at a minimum, the adjustments to the six epoch state parameters, and may include adjustments to various dynamic, geometric, and clock parameters as well. Equation (5) is called the regression equation and A is the matrix of regression coefficients.

A detailed discussion of the construction of A is beyond the scope of this chapter, but a simple overview is in order. An element a_{ij} of A is given by the following equation

$$a_{ij} = \frac{\partial z_i}{\partial x_j} \tag{6}$$

where, for simplicity, z_i now represents the differential element δz_i. This partial derivative relates an observation z_i at one time to state parameter x_j at a possibly remote reference time. The A matrix, thus, contains the state transition information from the reference epoch to all times in the data arc and must, therefore, embody the dynamical constraint of Eq. (2). To compute the a_{ij}, we first write

$$\frac{\partial z_i}{\partial x_j} = \frac{\partial z_i}{\partial \boldsymbol{x}_{ci}} \frac{\partial \boldsymbol{x}_{ci}}{\partial x_j} \tag{7}$$

where $\boldsymbol{x}_{ci}$ represents the satellite state at the time of observation z_i. This explicitly introduces the current state $\boldsymbol{x}_{ci}$ and its relation to both the current observation z_i and the state variables x_j. The partial $\partial z_i/\partial \boldsymbol{x}_{ci}$ contains no dynamical information and can be computed directly. The partial $\partial \boldsymbol{x}_{ci}/\partial x_j$ relates the satellite state at the observation time to the epoch state and, thus, embodies the dynamical constraint. To determine that partial, we differentiate the equation of motion (3) with respect to the epoch state parameters, producing a set of linear second-order differential equations in $\partial \boldsymbol{x}_{ci}/\partial x_j$. These *variational equations* are then integrated numerically to obtain the partial derivative and, thus, the final regression coefficients.

The well-known least squares solution to the regression equation (5) is given by

$$\hat{\boldsymbol{x}} = (A^T R_n^{-1} A)^{-1} A^T R_n^{-1} \boldsymbol{z} \tag{8}$$

where

$$R_n = E[\boldsymbol{n}\,\boldsymbol{n}^T] \tag{9}$$

is the covariance matrix associated with the measurement noise vector $\boldsymbol{n}$. This is known as the batch least squares solution because it requires that all observations over a data arc be collected and processed as a batch. In practice, when many parameters are estimated, Eq. (8) will require large matrix inversions, which can cause numerical instability. Most orbit estimators today employ more stable techniques.

C. Kalman Filter Formulation

A spaceborne GPS user may require a continuous real-time state solution more accurate than point positioning can provide. Although filtering is needed to achieve this, a batch solution is generally inappropriate because it may require a long accumulation of measurements and a large amount of computation at once. In such cases, a sequential estimator is called for, a popular example of which is the Kalman filter.

A sequential filter continually updates the current state estimate with each new measurement. The computation needed for each update is small compared with that for a full batch solution (although for a properly formulated filter, the computation required for many hours of updates is comparable to that for the same size batch solution); hence, an onboard processor can maintain the solution in real time. It should be noted that the sequential current state estimate employs only data from the past up to the present, whereas a batch filter may estimate a state with data from both before and after an epoch. In non-real-time uses, the final sequential state estimate can be mapped to all times in the data arc, just as in a batch solution, to achieve an equivalent result.

The conventional Kalman filter is formulated in discrete time recursion relations. Suppose the filter has produced a state estimate $\hat{x}_i$ at time t_i (using data up through time t_i) and that the estimated covariance matrix for $\hat{x}_i$ is $\hat{P}_i$. The state solution $\hat{x}_{i+1}$ at time t_{i+1} is derived in two steps: 1) the *time update,* in which a predicted or a priori solution $\tilde{x}_{i+1}$ and covariance matrix $\tilde{P}_{i+1}$ are generated from their estimated values at time t_i, with no new data yet added; and 2) the *measurement update,* in which the new estimates $\hat{x}_{i+1}$ and $\hat{P}_{i+1}$ are generated from the data at time t_{i+1}, as corrections to the predicted values.

The time update is given by

$$\tilde{x}_{i+1} = \Phi_i \hat{x}_i \tag{10}$$

and

$$\tilde{P}_{i+1} = \Phi_i \hat{P}_i \Phi_i^T \tag{11}$$

where Φ_i is the transition matrix derived from the equation of motion (or other appropriate transition models) relating the state at t_i to the state at t_{i+1}. The measurement update is then

$$\hat{x}_{i+1} = \tilde{x}_{i+1} + G_{i+1}(z_{i+1} - A_{i+1}\tilde{x}_{i+1}) \tag{12}$$

and

$$\hat{P}_{i+1} = \tilde{P}_{i+1} - G_{i+1} A_{i+1} \tilde{P}_{i+1} \tag{13}$$

where z_i is the measurement vector at time t_i, A_i is the matrix of measurement partials with respect to x_i, G_i is the so-called Kalman gain, given by

$$G_i = \tilde{P}_i A_i^T (A_i \tilde{P}_i A_i^T + R_{ni})^{-1} \tag{14}$$

and R_{ni} is the error covariance of the measurement vector z_i. (In some cases, for example when onboard computing is limited, a suboptimal *fixed gain* filter is employed, in which G is predetermined.) Note from Eq. (14) that, like the batch formulation, the conventional current state Kalman filter involves matrix inversion. Various alternative approaches have been devised that employ *pseudo-epoch state* factorized formulations.[23,24] These avoid matrix inversion by factoring P into either upper triangular and diagonal matrices (*U-D* formulation) or its square root matrices (square root information filter formulation). Factorized filters have been incorporated into several of NASA's high-performance orbit determination systems. For more on these techniques see Ref. 41.

Comparison of the batch and sequential formulations reveals that the latter is simply a recursive equivalent of the former. For a given data arc, the final sequential and batch solutions, when mapped to the same epoch, will be identical. As presented here, both are dynamical formulations that depend fundamentally on physical force models to produce the solution trajectory. It is worthwhile at this point to examine the principal errors that arise in the dynamic state solutions.

D. Dynamic Orbit Error

The typical accuracy of instantaneous point positioning is 10–20 m for the PPS user; the major error contributors are GPS orbit and clock error and pseudorange measurement error. Filtering reduces the position error by smoothing measurement error against an orbit model over the fitting arc. Meter-level random errors may readily be reduced to decimeters or below. At the same time, key systematic errors—GPS orbits and clocks, multipath—may be largely uncorrelated with the low orbiter dynamics and, therefore, attenuated in the solution. (Errors that correlate strongly with the orbit, such as once-per-orbit ionospheric effects, may be amplified.) Dynamic filtering can lower real-time orbit error to a few meters for the PPS user and to 20 m or better for the SPS user.

This improvement does not come without a cost. As the filter smooths measurement error, it introduces dynamic model error. Because force models are imperfect, the model trajectory will be imperfect. Model adjustments made during the solution may offer only partial improvement. Any remaining model errors will appear directly in the orbit solution. Gravity and drag model errors are often dominant, and both increase rapidly as the satellite altitude is reduced. Thus, accurate dynamic orbit estimation becomes problematical at low altitudes. For example, the motion of Lageos, a dense, inert sphere at about 6000 km altitude, can be modeled to within a few centimeters over periods of weeks; the motion of TOPEX/Poseidon, a larger vehicle at 1336 km, to about 10 cm over 10 days; the motion of SEASAT, at 800 km, to one or two meters over one day; and the motion of the Shuttle, at 300–400 km, to roughly 10 m over an orbit. At very low altitudes, dynamic filtering may offer little advantage over simple point positioning.

Model errors often reveal themselves in the postfit residuals; that is, they create systematic discrepancies between the actual measurements and theoretical measurements derived from the solution trajectory. Imagine a case in which a force varies randomly from one time step to the next and is, therefore, unpredictable, but can be observed in the postfit residuals. At some level, a number of forces (drag, gravity anomalies) can seem to behave that way. What is needed, then, is a means of extracting information in the residuals to recover unmodeled motion. The Kalman filter provides such a means in the form of process noise modeling.

E. Kalman Filter with Process Noise

To observe unmodeled motion, we model the time-varying satellite force as the sum of a deterministic component (our standard dynamic model) and a stochastic component. The latter is often called a process noise model. Augmenting a Kalman filter with a process noise model is a way of telling the filter that the state transition information in Φ is incomplete—that there is another component that the filter cannot predict, but that it can try to observe in the data and estimate at each time step.

In the context of orbit determination, this means that at each time step, in addition to applying the standard dynamic updates, the filter will examine the discrepancy between the dynamic state estimate and the apparent state as indicated geometrically by the measurements. From that discrepancy, it will estimate a local correction to the dynamic model, valid only over the update interval (t_{i-1},

t_i). When added to the dynamic model, that correction will reduce the disagreement between the observations and the solution trajectory at time t_i. As it proceeds through the data, the filter will generate a sequence of local force model corrections, one at each update time, bringing the solution trajectory into better agreement with the observations. That may be good or bad, depending on the quality of the observations and the accuracy of the models. We must, therefore, take care to hinder the local corrections from chasing after bad measurements.

The process noise model can take many forms, and various constraints may be applied to limit the freedom of each new correction to depart from the dynamic model or from the previous correction. The stochastic correction may be introduced by augmenting the state vector $\boldsymbol{x}_i$ with a vector $\boldsymbol{p}_i$ representing the local force model adjustment to be estimated at time t_i. For this discussion, we let $\boldsymbol{p}_i = (p_{i1}, p_{i2}, p_{i3})^T$ denote a three-dimensional force that is constant over the interval (t_{i-1}, t_i) and zero elsewhere. This force will be estimated to minimize the discrepancy between the dynamic solution update and the observations at time t_i. The augmented state vector $\boldsymbol{X}$ is given by the following equation

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{p} \end{bmatrix} \tag{15}$$

An effective realization of the process noise sequential filter used extensively by NASA in orbit estimation is given below.[23,26]

The time update requires an important modification. We have

$$\tilde{\boldsymbol{X}}_{i+1} = \Phi_i \hat{\boldsymbol{X}}_i \tag{16}$$

and

$$\tilde{P}_{i+1} = \Phi_i \hat{P}_i \Phi_i^T + B\,Q_i B^T \tag{17}$$

where now

$$\Phi_i = \begin{bmatrix} \Phi_x(i+1, i) & \Phi_{xp}(i+1, i) \\ 0 & M_i \end{bmatrix} \tag{18}$$

$$B = \begin{bmatrix} 0 \\ I_p \end{bmatrix} \tag{19}$$

Φ_x is the dynamic transition matrix of Eq. (10); $\Phi_{xp}(i+1, i)$ is the transition matrix relating $\hat{\boldsymbol{x}}_{i+1}$ to the process noise parameters $\boldsymbol{p}_i$; M_i is a 3 × 3 diagonal matrix with the jth element

$$m_j = \exp\,[-(t_{i+1} - t_i)/\tau_j] \tag{20}$$

Q_i is a diagonal covariance matrix associated with a white noise process, with the jth element

$$q_j = (1 - m_j^2)\,\sigma_j^2 \tag{21}$$

and $\boldsymbol{I}_p$ is a 3 × 3 identity matrix. The measurement update equations are identical to Eqs. (12–14), except that now we use the augmented state vector $\boldsymbol{X}$ and its associated covariance matrix P.

This is a first-order Gauss-Markov process noise model. (For some other possibilities, see Ref. 27.) Note that M_i is the transition matrix for the process noise parameters, and that the transition is in the form of a decaying exponential correlation. The time constant τ_i in Eq. (20) can be chosen to reflect the correlation in the dynamic model error (and thus in the desired correction) over one update interval. If τ_i is much smaller than the update interval, then m_i is small; the model error, therefore, is regarded as uncorrelated from batch to batch, and this becomes a white noise error model. There is one other selectable parameter, the steady state variance σ_i^2. Through Eq. (21), σ_i^2 scales the batch-to-batch variance q_i, which further constrains the correction. In the case of a white noise model, this constrains each independent correction with respect to the dynamic model, with no dependence on the previous correction. If $\sigma_i = 0$, the local force correction is constrained to zero and the conventional dynamic solution is obtained. In summary, the constraint is determined by τ_i (through m_i) and σ_i^2 as they combine through Eq. (21) to form the weighting matrix elements q_i.

Stochastic force models introduce an additional complication for non-real-time applications in which an optimal solution over an entire data arc is desired. It is no longer sufficient simply to map the final state solution to other times by means of the final dynamical models. The local force corrections have been determined with data only up to the times they occur and, thus, have not benefitted from later measurements. To complete the estimates of the local forces it is necessary to filter the data in the reverse direction as well, a process called smoothing, before mapping to all time points. The combined estimator is known as a filter/smoother.

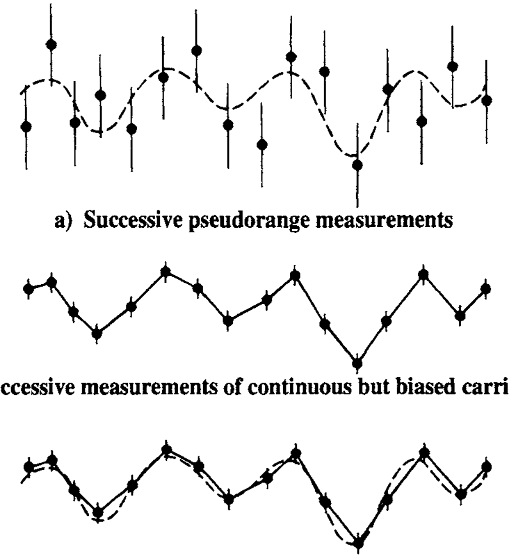
With conventional (sparse) tracking data we must be careful when employing process noise model corrections. The data acquired at any one time are often weak (or nonexistent), and insufficient by themselves to determine position. A relaxed constraint on the process noise estimate may result in a large and erroneous adjustment to the state, or may cause the solution to fail. Care must be taken to constrain the corrections within the observability limits of the data. This has traditionally meant relatively long correlation times and tight sigmas.

III. Orbit Estimation with GPS

We are now in a position to examine the powerful advantages GPS brings to estimating satellite orbits. First we look at a simple geometric technique that can improve point-positioning accuracy without dynamic filtering by combining the continuous carrier phase and pseudorange observables.

A. Carrier-Pseudorange Bias Estimation

When pseudorange and continuous carrier phase are brought together, point position accuracy can be improved by exploiting the absolute pseudorange information to estimate the bias in carrier phase. The bias is estimated simply by averaging the difference between phase and pseudorange for as long as the phase remains continuous (and the bias remains constant). This converts biased phase to a precise pseudorange with a small residual bias, preserving the detailed information on range change in carrier phase. The concept is illustrated in Figs.



a) Successive pseudorange measurements

b) Successive measurements of continuous but biased carrier phase

c) Absolute phase derived by adjusting mean of b to mean of a

Fig. 1 In carrier-pseudorange bias estimation, the phase bias is estimated by averaging the difference between continuous carrier phase and independent pseudoranges. The result retains the precision and time resolution of carrier phase, while reducing the bias to a fraction of the pseudorange error.

1a–1c. A sequence of N independent pseudorange measurements $\hat{x}_k$ is shown in Fig. 1a. The true time-varying pseudorange is represented by the dashed line. If x_k is the true pseudorange at time t_k, we can write the following:

$$\hat{x}_k = x_k + n_k \tag{22}$$

where, for simplicity, we assume n_k is a white noise process with standard deviation σ_n. Figure 1b shows the record of pseudorange change obtained by tracking carrier phase over the same arc. This can be regarded as a series of pseudorange measurements, $\bar{x}_k$, having a much smaller random error and a common bias. Thus, we can write

$$\bar{x}_k = x_k + b + e_k \tag{23}$$

where b is the bias and e_k is a white noise process with standard deviation σ_e. We estimate the bias b by averaging the difference between the $\bar{x}_k$ and $\hat{x}_k$

$$\hat{b} = \frac{1}{N}\sum_{k=1}^{N} \bar{x}_k - \hat{x}_k \tag{24}$$

or

$$\hat{b} = b + \frac{1}{N}\sum_{k=1}^{N} e_k - n_k \tag{25}$$

Because σ_e is typically 100 times smaller than σ_n, the approximate component error on the bias estimate is the following:

$$\sigma_{\hat{b}} \cong \frac{\sigma_n}{\sqrt{N}} \tag{26}$$

Thus, meter-level random noise on 1-s pseudorange data can give a decimeter-level bias estimate within 2 min. Subtracting Eq. (25) from Eq. (23) eliminates the bias in the phase measurement to give a precise record of absolute pseudorange. As shown in Fig. 1c, the corrected phase measurements sit close to, and have nearly the exact shape of, the true pseudorange sequence. The corrected phase measurements will have an approximate error

$$\sigma_x = (\sigma_{\hat{b}}^2 + \sigma_e^2)^{1/2} \tag{27}$$

where $\sigma_{\hat{b}}$ represents the residual bias common to all phase points and σ_e is the point-to-point random error. A sequence of position solutions derived from the corrected phases will have the precision of a pure carrier solution, with a bias that is a fraction of the typical point position error.

This technique, which was first proposed in Ref. 28, can be readily generalized to provide real-time recursive estimation of the position of an unpredictably moving vehicle. Consider a receiver that produces an instantaneous point position solution $\hat{x}_k$ at time t_k, and a position change solution $\Delta\hat{x}_k$ obtained by continuously tracking carrier phase from t_{k-1} to t_k. An estimate $\hat{X}_{n+1}$ of the position at time t_{n+1} is given by the following

$$\hat{X}_{n+1} = \frac{n}{n+1}(\hat{X}_n + \Delta\hat{x}_{n+1}) + \frac{1}{n+1}\hat{x}_{n+1} \tag{28}$$

Note that this is a variation on the recursive formula for a simple average

$$\overline{P}_{n+1} = \frac{n}{n+1}\overline{P}_n + \frac{1}{n-1}P_{n+1} \tag{29}$$

The position change information $\Delta\hat{x}_{n+1}$ maps the current position estimate $\hat{X}_n$ forward to the next time point for averaging with the point position $\hat{x}_{n+1}$ computed at that time. Carrier phase, in effect, inertially aids the sequential averaging of point position solutions to refine the phase bias estimate. The procedure can be tuned by weighting each $\hat{x}_k$ by its inverse covariance.

A principal virtue of this technique is its extreme simplicity. A filter to track unpredictable motion (or the relative positions of multiple vehicles) can be realized in a relatively few lines of code. It is, however, suboptimal. Correlation between the $\Delta\hat{x}_k$ is not properly accounted for, and it does not fully exploit the information in the carrier phase. More refined strategies are presented in Ref. 29. Another drawback is its exclusion of external information about platform dynamics. The solution becomes vulnerable to outages that might easily be bridged with simple dynamic models. These weaknesses are remedied in a more robust technique that employs the Kalman filter formalism.

B. Kinematic Orbit Determination

When a Kalman filter is applied to GPS data from a low orbiter, the full advantage of continuous three-dimensional coverage may not be realized without an aggressive use of process noise corrections. If we assume a full GPS constellation, a flight receiver having six or eight parallel channels, and a relatively wide field of view, strong instantaneous observing geometry is assured. Inclusion of continuous carrier phase data vastly increases the potential precision of the estimates. It then becomes possible to relax or eliminate constraints on the process noise force corrections and track the true motion of the vehicle with great precision.

The concept is illustrated in Fig. 2. The dashed curve represents the irregular path of a low orbiter subject to varying forces. With GPS data collected by the orbiter, we can execute a traditional dynamic orbit solution to produce the smooth orbit estimate shown by the solid line. This leaves a set of (possibly large) postfit residuals. Because GPS provides continuous three-dimensional coverage, the postfit residuals at each time point suffice to reconstruct the observed satellite position (its departure from the dynamic solution) by purely geometric means. The observed trajectory can then be recovered by adding the geometrically determined correction to the dynamic solution at each time point. Force model error, reflected in the initial postfit residuals, is thereby eliminated.

This can be thought of as two distinct steps. First, a conventional dynamic solution produces a reference trajectory and postfit residuals; the residual path is then constructed geometrically, point by point, and added to the dynamic solution. In practice, this can be done in one estimation step in a Kalman filter with process noise. The estimated process noise parameters $\boldsymbol{p}$ in Eq. (15) can provide the geometric corrections to the dynamic solution. In ordinary tracking applications, those parameters would be tightly constrained and geometric information only weakly expressed. But the full observability offered by GPS allows all constraint to be removed. The correlation time τ_i can be set to zero (white noise model) and the steady-state variance σ_i^2 to a large value. The filter will then estimate a three-dimensional force correction for each interval (and a corresponding change in the current state) to account for the geometric discrepancy between the measurements and the dynamic solution. This is called nondynamic or *kinematic* orbit determination, although both terms are somewhat misleading because the technique builds on an underlying dynamic formulation.

As we see in more detail later, the kinematic solution can be carried out with pseudorange data alone, with carrier phase data alone, or with the two in

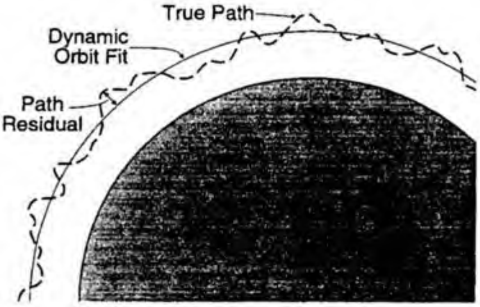


Fig. 2 The kinematic orbit determination technique effectively reconstructs the observed trajectory from the residuals of a dynamic orbit solution.

combination. Observe, however, that as the dynamic constraint is relaxed to allow the geometric correction, the effect of measurement error increases. Instead of being smoothed against the dynamic model, single-point measurement error is fully expressed in the geometric correction. Thus, when pseudorange alone is used, the kinematic solution becomes a series of point positions with full pseudorange noise. For precise applications, continuous carrier phase is essential.

C. Reduced Dynamic Orbit Determination

Because the kinematic correction is geometric, it is vulnerable to weak geometry. Momentary data outages or large position dilutions of precision (PDOPs) will cause the error to grow or the solution to fail. The kinematic solution, moreover, makes little use of dynamic information—it is an empirical result constructed from the measurements. Often, however, dynamic information is at hand, which, if properly treated, can improve the result. When geometry weakens or fails, dynamic information can then carry the solution with little loss of accuracy.

We can achieve a balance of dynamic and geometric information in the orbit solution by imposing a judicious constraint on the process noise parameters. In an optimal solution (under the assumption of a Gauss-Markov process noise model), the time constant τ_i will reflect the actual correlation time of dynamic model errors, and the steady-state variance σ_i^2 the actual error in the dynamic model. The geometric corrections will not be free to follow the measurements wherever they lead, but will be bound by the constraint to the dynamic model. Relative weight will, in fact, shift back and forth between dynamic and geometric information as observing strength varies. When geometry is weak, the process noise constraint will hold the correction close to the dynamic solution; if there are no observations at all, no correction can be computed and the dynamic solution is produced. This optimized technique is known as *reduced dynamic* orbit determination.[42]

Another interpretation is given in Fig. 3, which illustrates the relative significance of random and systematic error in the solution trajectory. In the dynamic solution, random error is minimized (because the fewest parameters are adjusted), while dynamic error is fully expressed. This is reversed in the kinematic solution as many parameters are adjusted, amplifying the effect of data noise while absorbing dynamic error. The reduced dynamic solution seeks the optimal balance to minimize overall error.

This raises the question of how we choose the process noise weighting. Often there is some prior knowledge of the quality of the force models in use and the consequent position error expected. Computer simulations or covariance analysis can then suggest a reasonable a priori weighting. When real data become available, a variety of strategies for tuning the reduced dynamic constraints become possible. One approach is to observe the magnitude of the process noise corrections; if they approach the constraints, the constraints should be relaxed; if they fall well short, the constraints can be tightened. Another technique is to compare orbit solutions on short overlapping segments and tune the constraints to minimize the discrepancy.

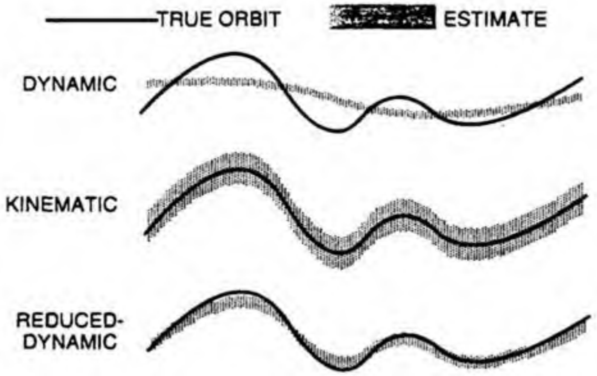


Fig. 3 The purely dynamic orbit solution minimizes the contribution of random error, while dynamic model error is fully expressed; this is reversed in the kinematic solution. The reduced dynamic solution yields an intermediate level of each error and can minimize overall error.

The reduced dynamic technique is one realization of the concept depicted in Fig. 2; others are possible. For example, we might compute position rather than force corrections. An approach along those lines proposed in Ref. 30 has certain advantages for gravity recovery. Force corrections, however, directly augment the dynamic model and have the virtue that, although discontinuous (piecewise constant) themselves, they yield a continuous trajectory when integrated.

D. Orbit Improvement by Physical Model Adjustment

The reduced dynamic solution employs local corrections to dynamic models. Often it is more efficient to adjust physical model parameters; fewer adjustments may be needed and data strength preserved. Adjustment of drag and radiation pressure terms, for example, is common. Particularly attractive with GPS tracking data is *gravity tuning,* or adjustment of gravity field coefficients. The geopotential is commonly represented as a spherical harmonic expansion containing anywhere from a few terms to a few thousand terms, depending on the fidelity required. Each gravity harmonic is a global function representing a permanent component of the geopotential.

Many geopotential models are derived from satellite tracking data, which are often sparse in some regions. GPS, however, leaves no coverage gaps. Because a polar orbiter overflies the entire globe, GPS tracking of such a satellite can enable improvement of the full global model. That will in turn reduce the dynamic model error and permit tighter constraints on the process noise models in subsequent orbit solutions. Gravity tuning has elements in common with reduced dynamic orbit estimation. Both techniques adjust a large number of force parameters to bring the solution trajectory into closer agreement with the data. Where gravity is the dominant model error, gravity tuning is a desirable first step, because it yields a permanent model improvement.

IV. Direct Orbit Determination with GPS

Sophisticated estimation strategies may be of little value in direct GPS-based orbit determination, where only the onboard observables and broadcast data are used in the solution. Although measurement noise can be reduced to centimeters by filtering, final user orbit error will be dominated by GPS ephemeris and clock error (with possibly large contributions from the ionosphere and selective availability) at a level of meters to tens of meters. Evaluating the expected accuracy of the filtered orbit solution is not always straightforward.

Consider the batch least squares dynamic solution of Eqs. (8) and (9). It is easily shown that the error covariance P_x on the estimate $\hat{\boldsymbol{x}}$ is given by the following

$$P_x = (A^T R_n^{-1} A)^{-1} \tag{30}$$

This is the formal error attributable to the random measurement noise vector $\boldsymbol{n}$ (Eq. 5), sometimes called the *commission* error. It does not take into account other errors present in the solution, such as those attributable to GPS orbit and clock errors, sometimes called *omission* errors. To examine the effect of such errors, we can include the relevant parameters and their relation to the observations explicitly in the regression Eq. (5) by writing the following equation

$$\delta z = A\boldsymbol{x} + B\boldsymbol{y} + \boldsymbol{n} \tag{31}$$

where $\boldsymbol{y}$ is the vector of omission error parameters and B is a matrix of partial derivatives of the observations δz with respect to $\boldsymbol{y}$. When the solution given by Eq. (8) is applied to Eq. (31) we have the following

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + (A^T R_n^{-1} A)^{-1} A^T R_n^{-1} B\,\boldsymbol{y} + \tilde{\boldsymbol{n}} \tag{32}$$

where $\tilde{\boldsymbol{n}}$ is the transformed random measurement noise. The long expression multiplying $\boldsymbol{y}$ in Eq. (32) describes the response of the estimate $\hat{\boldsymbol{x}}$ to the error parameters $\boldsymbol{y}$, and is called the sensitivity matrix S

$$S = (A^T R_n^{-1} A)^{-1} A^T R_n^{-1} B \tag{33}$$

The total error covariance P_{tot} of the estimated vector $\hat{\boldsymbol{x}}$ is given by the following

$$P_{\text{tot}} = P_x + S\,P_{\text{om}}\,S^T \tag{34}$$

where P_{om}, the a priori covariance matrix for the omission errors, must be derived through careful analysis of those errors.

Because many omission errors are physically unrelated and can be regarded as uncorrelated, P_{om} can often be (and almost invariably is) set up as a diagonal matrix. The errors on the elements of a dynamic satellite state solution, however, are strongly correlated. (Note, for example, that there is a direct relationship between satellite altitude and in-track velocity; in a dynamic solution, an error in one will appear as a compensating error in the other.) A diagonal covariance matrix is therefore inadequate to assess the effect of GPS ephemeris error on a dynamic user orbit solution and, in fact, can be shown to give a highly pessimistic estimate of the error that actually arises. To evaluate the effect of GPS orbit error on a dynamic user solution, a full covariance matrix is needed for the GPS state parameters. One way to obtain such a matrix is to simulate the GPS orbit determination process as it is carried out with ground data to produce the GPS orbits available to the user.

One such study[31] showed that errors in the GPS orbits were attenuated by roughly a factor of two in the dynamic solution for an orbiter at 1300 km. That is, GPS orbit errors of 1 m resulted in errors of about 0.5 m in the user orbit. (When a diagonal GPS covariance matrix was used, this error was overestimated by a factor of about 20.) Because the satellite was at an altitude of 1300 km, model error was small and the full benefit of dynamic filtering could be gained. With a typical GPS broadcast ephemeris error of 5 m, we could expect to achieve a real-time orbit accuracy of 2–3 m for such a satellite, limited by the GPS orbit error.

At lower altitudes, dynamic model error grows. At 500 or 600 km dynamic error may equal GPS ephemeris error. Below 500 km, dynamic error will dominate, and the optimal filter will deweight dynamics. For the Space Shuttle at 300 km, the optimal solution will be nearly kinematic, and position error, dominated by GPS orbit error, will be little better than with simple point positioning. Because GPS orbit errors change slowly, the direct kinematic error will be highly correlated from one second to the next.

The most accurate direct orbit solutions are therefore obtained by dynamic filtering for satellites above about 800 km (and below 3000 km), with the accuracy limited by GPS orbit error. The best GPS orbits produced today (available typically several days after the fact) are accurate to better than 1 m. In principle, such accuracies can be achieved nearly in real time. Moreover, accuracies of 1–2 m can be reached for GPS orbits and clocks predicted several hours into the future[25] and, thus, available for true real time use. If dynamic filtering reduces the resulting user error by a factor of two, real-time dynamic tracking could be made accurate to about 1 m above 800 km. At the lowest altitudes, where the kinematic solution is optimum, accuracy could reach a few meters. For further improvement at all altitudes we must reduce GPS orbit error.

V. Precise Orbit Determination with Global Positioning Systems

A few classes of mission require orbit accuracies ranging from 1 m (land altimetry, precise imaging) down to a few centimeters (ocean altimetry, gravity field modeling). For that we must turn to the techniques of differential GPS. As it has been developed for precise orbit determination, differential GPS is intended for non-real-time applications and differs considerably from the real-time differential techniques used for regional navigation.

A. Global Differential Tracking

The fundamental concept is illustrated in Fig. 4. In addition to the flight receiver, a network of reference receivers around the world continuously tracks all GPS satellites in view. The flight receiver and at least one ground receiver must share common visibility of several GPS satellites at all times. Only about six well-distributed ground sites are needed to ensure this, although in recent experiments with TOPEX/Poseidon a dozen or more have been used. Several ground receivers may be at *fiducial* sites—sites with accurately known absolute

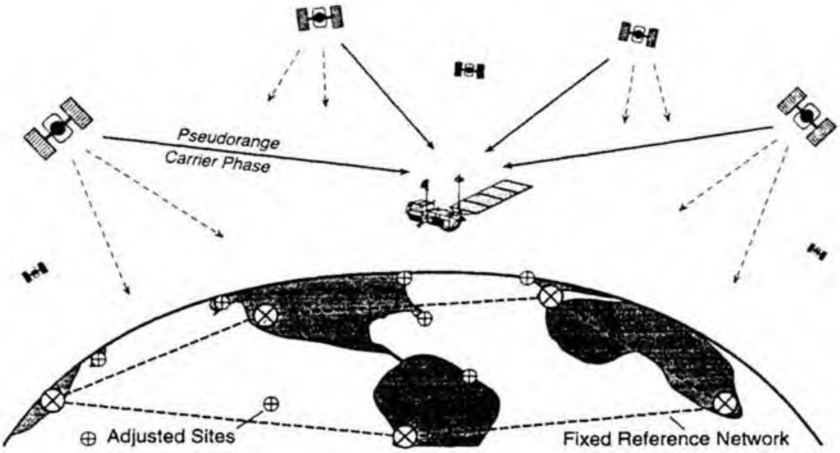


Fig. 4 In precise orbit determination with differential GPS, user and ground observations of GPS are combined to determine user, GPS, and some ground positions with respect to a subset of ground reference or "fiducial" sites.

positions that will be held fixed during the solution. The best current fiducial sites [those of the International Terrestrial Reference Frame (ITRF) maintained by the International Earth Rotation Service (IERS) in Paris] are known relatively to 1–2 cm, and absolutely (with respect to the geocenter) to better than 3 cm.

Pseudorange and carrier phase data from the flight and ground receivers are processed together to produce a single grand solution. The solution strategy can vary greatly in detail, but typically includes estimation of all GPS satellite orbits, the user orbit, transmitter and receiver clock offsets, all carrier phase biases, nonfiducial ground site positions, atmospheric propagation delays at all ground sites, and such satellite force parameters as atmospheric drag and solar radiation pressure. Data arc lengths may range from a few hours to many days. Because only the fiducial sites are held fixed, they establish the reference frame in which all other positions are determined.

One variation permits *all* ground sites to be adjusted within a moderate constraint, typically 10–1000 m on each site. This severs the tie to a predetermined reference frame and allows the entire solution to rotate within the limits of the overall constraint. The solution is then mapped into a chosen reference frame (such as ITRF) through a seven-parameter transformation (translation, rotation, and scale), which minimizes the three-dimensional rms difference between all ground site solutions and their values in the chosen frame. This removes dependence on a particular subset of sites to define the reference frame and reduces reference station error in the total error budget. A less powerful variation processes the ground and user data separately. The ground data first determine accurate GPS orbits and clock offsets, which are then applied in a direct user solution. This is less powerful than a true simultaneous solution, but may offer greater flexibility and convenience.

Although global differential tracking constitutes a major logistical departure from direct tracking, the basic filter equations needed to carry it out, Eqs. (16–21), remain unchanged. What changes is the definition of the estimated state vector X. To the user state and other adjusted parameters we now append state elements for all GPS satellites, clock offsets for all transmitters and receivers, ground site positions, atmospheric delays, and so on. The matrices of measurement partials and a priori covariance are correspondingly augmented, and the solution becomes more computationally demanding. It is worth examining in more detail how some of the key parameters are treated.

B. Fine Points of the Global Solution

When carrier phase data are used in a grand solution, either alone or together with pseudorange, the effective data noise (random measurement error) is typically below 1 cm. This can be seen in the postfit residuals of global geodetic solutions, which for the combined dual-frequency phase observable are typically 3–6 mm. As revealed in numerous covariance studies [Eq. (34)], random measurement error will contribute on the order of 2 cm to the user position error—somewhat higher for purely kinematic solutions and lower for purely dynamic solutions. In the grand solution, the major systematic model errors that plague the direct solution (GPS orbits and clocks) are reduced. Note, however, that if GPS satellite dynamics and clocks are poorly modeled, the GPS orbit and clock estimates will degrade and systematic errors will still arise in the user state solution. Fortunately, the high-altitude GPS satellite dynamics can be well modeled over 24 h, and standard dynamic GPS solutions generally suffice. For longer arcs, a weak stochastic adjustment of the GPS solar pressure parameter may be advantageous.

For clock solutions, we have several options. If high-quality atomic clocks are used in all receivers and transmitters, simple quadratic models might suffice over many hours. Because real clock behavior can be unpredictable, common practice is to allow for the worst by solving for all clock offsets independently at each time-step under a loose constraint. This is equivalent to modeling clock behavior as a white noise process with large variance, in analogy with our treatment of the process noise force parameters in the kinematic orbit solution. It is also similar to the popular practice of double differencing to eliminate clock parameters; however, when global data sets are used, as they must be for precise orbit determination, the white noise clock model is more powerful, as it retains more data.[32] Just as purely kinematic orbit determination fails to exploit known dynamics of the satellite, white noise clock models fail to exploit known (and perhaps smooth) clock behavior and, thus, must be regarded as a conservative strategy.

Computer simulations, covariance studies, and results with TOPEX/Poseidon have shown that the grand solution strategy can reduce user satellite position errors caused by GPS orbit and clock errors to a few centimeters. What, then, becomes the dominant error in the user solution? One candidate is the error in modeling atmospheric propagation delay at the ground sites, or, rather, the variable wet component of that delay. When standard seasonal models (supported by surface weather data) are used to calibrate the atmospheric delay, the error is typically 3–5 cm at zenith, which may translate into 2–10 cm of user state error, depending on the solution technique. This can be reduced by periodically solving for a zenith delay at each site. The most effective strategy yet developed is to model the zenith delay as a stochastic process (a random walk, for example) and adjust it at each time-step under a constraint derived from the observed power spectrum of atmospheric delay variation. Typical zenith delay accuracy with this technique is about 1 cm.

Finally, we note that each carrier phase observable contains an arbitrary bias corresponding to integer cycle ambiguities at each frequency. Those biases must be estimated (or eliminated by time differencing) when the phase observable is used. In precise ground-based geodesy, an effort is often made to determine the exact integer cycle ambiguities in the differential observables and then fix the biases at those values. Resolving ambiguities between an orbiter and ground sites is demanding and, when many hours of data are used, can be shown to contribute little to solution strength, because at that point data noise is not a dominant error. The differential strategies described here attempt no cycle ambiguity resolution, and instead treat each bias as a continuous variable.

C. Precise Orbit Determination Performance Analysis

It may now be evident that the general strategy for achieving high accuracy with GPS is to exploit the great strength of GPS data to observe and correct systematic errors that threaten to dominate. Just how the data will stand up to this demand depends on many details of system configuration and solution strategy. To illustrate those dependencies, we present the results of covariance studies for several real or proposed missions. All studies include both commission and omission errors in an attempt to arrive at realistic final error estimates.

The first example is taken from error studies conducted for TOPEX/Poseidon years in advance of its launch in August of 1992. TOPEX/Poseidon is a U.S.-French ocean altimetry mission flying at an altitude of 1336 km, where dynamic model errors are now well below 10 cm. The GPS configuration for TOPEX/Poseidon includes a six-channel (dual-frequency) flight receiver with a hemispherical field of view, and a six-site ground network. The assumptions, error model, and estimation strategy are summarized in Table 1. Note that a reference frame error of 5 cm per component for each of three fiducial sites was assumed, far greater than that error today.

Figure 5 shows the predicted rms altitude error for three solution strategies—dynamic, kinematic, and optimized reduced dynamic—as a function of the gravity model error. Because the kinematic solution eliminates dynamic error, it is independent of the gravity model. Its total error is divided almost equally between data noise and reference site error. The dynamic solution error depends strongly on gravity error and becomes limited by data noise and reference frame errors only when gravity and other dynamic errors are small. The reduced dynamic strategy surpasses both kinematic and dynamic—the latter even when the gravity error is zero, because other dynamic errors are still reduced.

Also shown in Fig. 5 are three circles representing actual results from TOPEX/Poseidon obtained during the first year of the mission. The circles give the rms altitude agreement over 30 days between purely dynamic solutions made with ground-based laser ranging and Doppler data, and the GPS reduced dynamic solutions. An rms agreement of about 6 cm was obtained with the final prelaunch

Table 1 **Error model for TOPEX/Poseidon orbit determination analysis**

System characteristics	
Orbit (circular):	1334 km, 66-deg inclination
Number of ground sites:	6 (including 3 fiducial sites)
Number of GPS satellites:	18
Flight antenna field of view:	Hemispherical
Flight receiver tracking capacity:	6 channels (*L1* and *L2*)
Data types:	*L1* and *L2* pseudorange *L1* and *L2* carrier phase
Data interval:	5 min
Smoothed data noise:	5-cm pseudorange 1-cm carrier phase
Adjusted parameters and a priori errors	
TOPEX/Poseidon epoch state:	1 km; 1 m/s, each component
GPS satellite states:	2 m; 0.2 mm/s, each component
Carrier phase biases:	10 km
GPS and receiver clock biases:	3 ms (modeled as white noise)
Non-fiducial ground locations:	20 cm each component
Fixed errors evaluated	
Fiducial site positions:	5 cm each component
GM of Earth uncertainty:	1 part in 10^8
Earth gravity error model:	0–100% GEM10–GEML2 (20 × 20)
Zenith atmospheric delay error:	1 cm (modeled as random walk)
Atmospheric drag error:	10% of total
Solar radiation pressure error:	10% of total

gravity model, known as JGM-1, which has a quality roughly in the center of the range shown. This improved to about 3.5 cm with the JGM-2 model, which had been tuned with laser and Doppler data by the Goddard Space Flight Center several months after launch.[33] The agreement improved further, to about 2.5 cm, when the JGM-2 model was tuned with GPS data by investigators at the University of Texas.[34] At this point, the rms altitude error resulting from gravity mismodeling is believed to be no more than 2 cm.

Past ocean altimetry missions have been plagued by what are known as geographically correlated orbit errors, that is, orbit solutions that are consistently biased in different geographic regions. Such errors can confound the construction of global circulation models from the altimetry data. Geographically correlated orbit errors are often a consequence of geographic biases in the gravity model, although other factors may also play a role. Studies reported in Ref. 35 showed that kinematic and reduced dynamic orbits, by reducing dependence on force models in general, can virtually eliminate any geographic correlation in orbit errors resulting from the gravity model. That result was confirmed with TOPEX/Poseidon. Laser/Doppler dynamic orbit solutions with JGM-1 and JGM-2 showed consistent geographic discrepancies from the GPS reduced dynamic solutions.

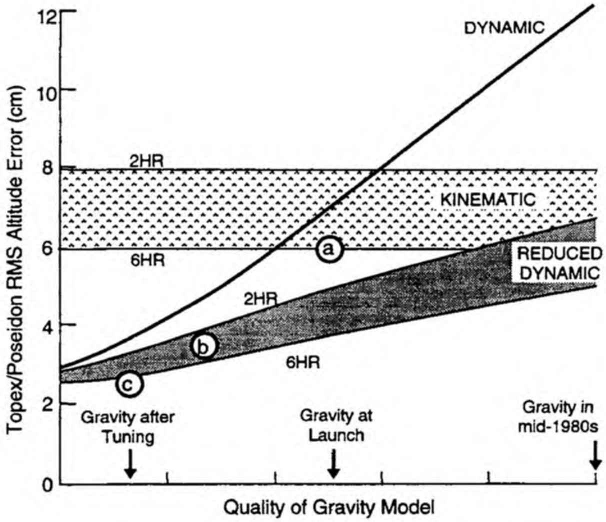


Fig. 5 Predicted rms altitude error for TOPEX/Poseidon as a function of the quality of the gravity model, for three different solution strategies (see Table 1). Circles show actual rms altitude agreement between GPS reduced dynamic and laser/Doppler dynamic solutions made with the prelaunch gravity model ⓐ, a laser/Doppler-tuned gravity model ⓑ, and a GPS-tuned gravity model ⓒ.

In later dynamic solutions made with the GPS-tuned gravity model, geographic discrepancies had all but vanished.

A second example is taken from the Earth Observing System, a suite of scientific Earth probes planned to fly at about 700 km beginning in the late 1990s. Because dynamic errors may grow large at that altitude, a purely kinematic analysis is presented. This time the reference site error is reduced to 3 cm per component. Other assumptions that differ from the TOPEX/Poseidon analysis are shown in Table 2. Figure 6 shows the resulting predicted altitude error as a function of data arc length for several different GPS data combinations. The data type called "carrier-quality pseudorange" is a fictitious pseudorange measurement having the precision of carrier phase, and serves to establish a performance bound.

Table 2 Changes from Table 1 for Earth Observing System kinematic orbit determination analysis

Orbit (circular):	705 km, 98-deg inclination
Number of GPS satellites:	24
Flight receiver tracking capacity:	All in view (within hemisphere)
Zenith atmospheric delay error:	Adjusted as random walk
Fiducial location error:	3 cm each component
Earth gravity error model:	100% GEM10–GEML2 (20 × 20)

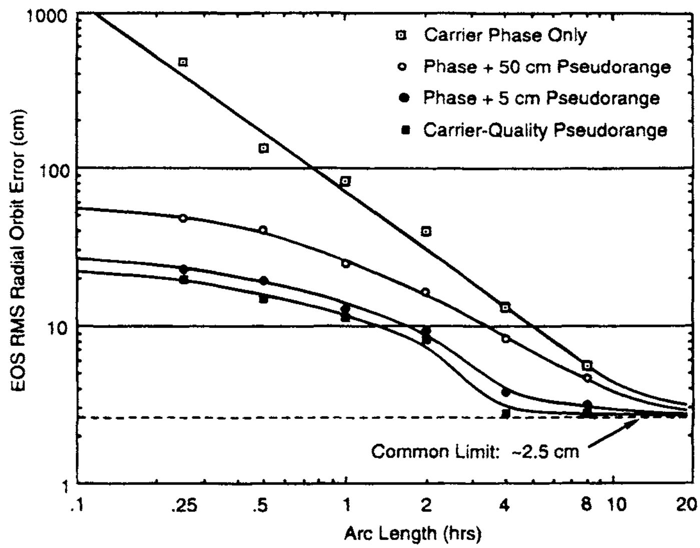


Fig. 6 Predicted rms altitude error for the Earth Observing System as a function of data arc length, for purely kinematic orbit determination and four data combinations. Key assumptions are shown in Table 2.

Figure 6 indicates that few-centimeter accuracy is possible for dynamically complex platforms, and that even the biased carrier phase observable used by itself can approach the performance of the strongest possible data type. It may seem surprising that the kinematic solution can succeed with carrier phase alone, considering that the grand solution must estimate phase and clock biases, which are nicely constrained by pseudorange data. But the dynamic core of the kinematic solution allows the biases to be reliably estimated, just as they are in an integrated Doppler dynamic solution. This illustrates a fundamental difference between the process noise Kalman filter formulation and the simple carrier-pseudorange bias estimation of Sec. III-A. The latter depends entirely on pseudorange to estimate the phase bias, whereas the former can recover the bias dynamically (while correcting the model kinematically) when range information is absent.

A third study explores the limits of kinematic performance with a stringent tracking challenge: the Space Shuttle at 300 km. For a given phase noise, kinematic tracking accuracy is limited largely by observing geometry, which we strengthen by assuming a full sky field of view (each Shuttle is equipped with GPS antennas top and bottom to permit this), a flight receiver able to track all satellites in view (typically 13–15), and 11 ground sites, with reference site error of 1.5 cm per component. Other assumptions are given in Table 3. As shown in Fig. 7, the limiting error in all components now approaches 1 cm, although in reality, dynamic errors in the GPS satellite orbit solutions might degrade this somewhat. This opens up new possibilities for near-Earth ocean altimetry and other precise Earth observations on platforms of opportunity, and for short-duration testing of precise instruments on the Space Shuttle.

Table 3 Changes from Table 1 for Shuttle kinematic orbit determination analysis

Orbit (circular):	300 km, 28-deg inclination
Number of GPS satellites:	24
Number of ground sites:	11 (including 3 fiducial sites)
Flight antenna field of view:	Full sky
Flight receiver tracking capacity:	All in view
Smoothed data noise:	5-cm pseudorange 5-mm carrier phase
Zenith atmospheric delay error:	Adjusted as random walk
Fiducial location error:	1.5 cm each component
Earth gravity error model:	50% GEM10–GEML2 (20 × 20)

D. Single-Frequency Precise Orbit Determination

The carrier-only kinematic solution is more than a curiosity. It enables accurate orbit determination with simple codeless receivers, bypassing the effects of anti-spoofing, and can be used to achieve fair orbit accuracy with single-frequency data as well.[44] In the examples thus far we have assumed dual-frequency elimination of ionospheric delay; but the ionosphere can also be removed by averaging the $L1$ phase and pseudorange observables. Consider these simplified expressions for the phase delay and group delay (pseudorange) observables

$$\tau_\phi \cong \tau - \frac{k \cdot \mathrm{TEC}}{f^2} + \mathrm{bias} + \epsilon_\phi \tag{35}$$

and

$$\tau_{\mathrm{grp}} \equiv \tau + \frac{k \cdot \mathrm{TEC}}{f^2} + \epsilon_{\mathrm{grp}} \tag{36}$$

where TEC is the total electron content along the raypath, f is the observing frequency, k is a constant, ϵ is the random measurement error; and τ is the common delay caused by geometry and factors other than the ionosphere. Note

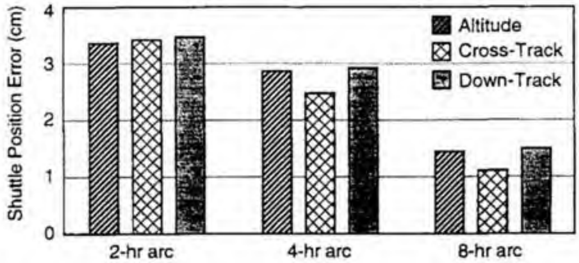


Fig. 7 Predicted kinematic tracking error for the Space Shuttle with a robust GPS observing system. Key assumptions are shown in Table 3.

that the ionosphere term is identical in both equations but appears with opposite sign. Forming the simple average of Eqs. (35) and (36) we obtain

$$\frac{\tau_\phi + \tau_{\mathrm{grp}}}{2} \cong \tau + \mathrm{bias}' + \frac{\epsilon_\phi + \epsilon_{\mathrm{grp}}}{2} \tag{37}$$

The ionosphere term is canceled, and the resulting observable has the form of the biased carrier phase delay [Eq. (35)]. This is sometimes called the GRAPHIC (Group And Phase Ionospheric Correction) observable. Because ϵ_{grp} is much greater than ϵ_ϕ the effective measurement error on Eq. (37) is half that of pseudorange. Note that the conventional dual-frequency correction *increases* raw data noise by a factor of three; thus, if single-frequency phase is 100 times more precise than pseudorange, dual-frequency phase will be only 17 times more precise than GRAPHIC data.

Modern receivers that employ 20 MHz C/A-code processing can recover C/A pseudorange with a precision of better than 50 cm in 1 s. The GRAPHIC observable reduces this by half. Smoothing over 60 s can bring the error below 10 cm. Figure 8 shows the predicted three-dimensional rms position error for the Shuttle at 300 km with three solution strategies: dual-frequency dynamic, GRAPHIC kinematic, and dual-frequency kinematic. Key assumptions are shown in Table 4. Note that drag and gravity errors make the dynamic solution worse than simple point positioning. The kinematic solutions improve orbit accuracy by two to three orders of magnitude, reaching about 2 cm per component with dual-frequency phase. The order-of-magnitude difference between the two kinematic cases is explained by the higher data noise on the GRAPHIC observable.

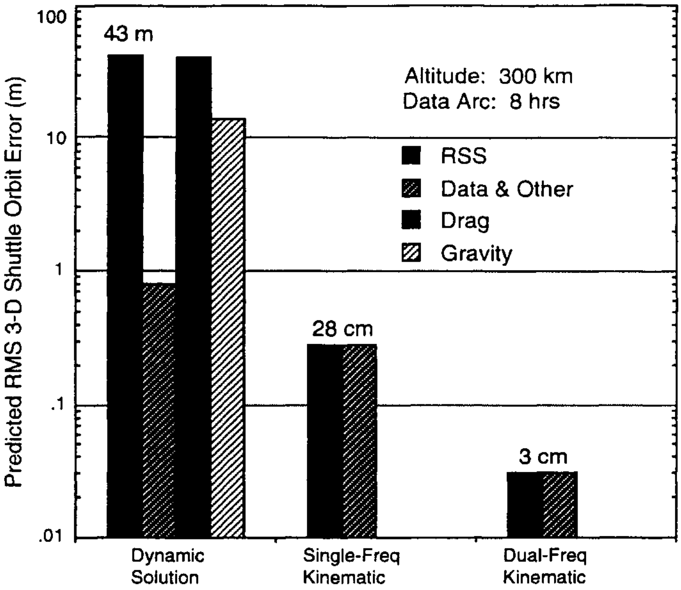


Fig. 8 Predicted three-dimensional error for the Space Shuttle at 300-km altitude, with three different solution strategies.

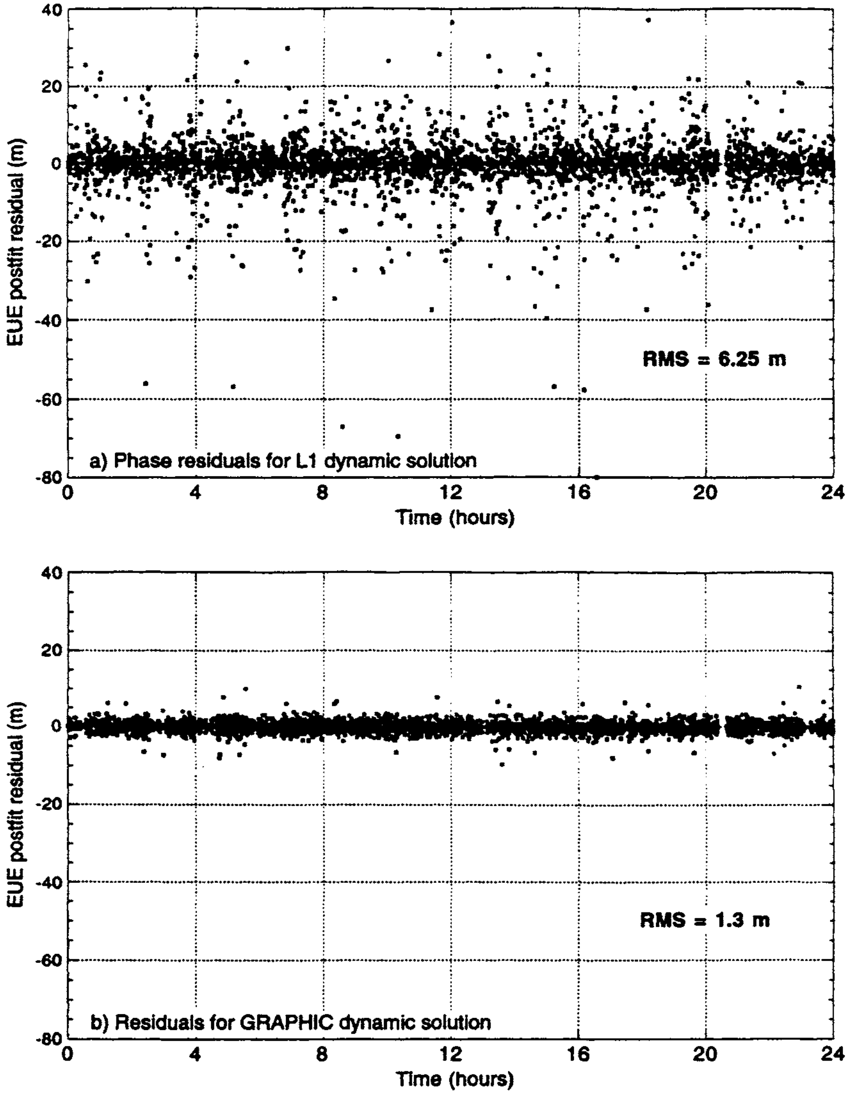


Fig. 9 Postfit residual plots for GPS-based EUE dynamic orbit solutions with a) single-frequency carrier phase data and b) the single-frequency ionosphere calibrated GRAPHIC observable.

Table 4 Key assumptions for Shuttle single-frequency kinematic orbit determination analysis

Orbit (circular):	300 km, 98-deg inclination
Number of GPS satellites:	24
Number of ground sites:	6 (including 3 fiducial sites)
Flight receiver tracking capacity:	All in view (within hemisphere)
Smoothed data noise:	10-cm pseudorange (single-frequency) 1-cm carrier phase
Zenith atmospheric delay error:	Adjusted as random walk
Fiducial location error:	3-cm each component
Earth gravity error model:	30% GEM10–GEML2 (20 × 20)
Atmospheric drag error:	10% of total

A variant of this technique was first described in Ref. 45. More recently, single-frequency ionosphere calibration was demonstrated on an Earth satellite in Ref. 36. The Extreme Ultraviolet Explorer (EUE), flying at about 500 km, is equipped with a 12-channel $L1$-only receiver and two oppositely directed antennas, providing a full sky field of view. Many GPS tracks acquired by EUE look down through the ionosphere, where the added delay can exceed 50 m. Figure 9 presents typical postfit residual plots for EUE GPS orbit solutions. In Fig. 9b, large ionospheric excursions are entirely absent. Direct comparison of orbit overlaps indicates an rms altitude error of less than 1 m in a differential reduced dynamic EUE solution with the single-frequency GRAPHIC observable.

E. Extension to Higher Altitude Satellites

Above about 3000 km, an orbiter begins to lose coverage from GPS. However, because dynamic model error can be small at high altitudes, the dynamic orbit solution can remain strong. By looking downward to catch the signal spillover from satellites on the other side of the Earth, an orbiter can exploit GPS from well above the GPS satellites themselves, out to geosynchronous altitude and beyond. Alternatively, high satellites can carry GPS-like beacons to be tracked from the ground, with the GPS satellites serving as reference points, a technique known as *inverted* GPS. Figure 10 (from Ref. 30) plots the average number of GPS satellites that can be tracked by a circular orbiter as a function of altitude for both upward- and downward-looking vehicles, where each is assumed to have a hemispherical field of view. Note that above about 2000 km, the down-looking user can track more. The figure also plots the average number of ground sites that can track a beacon on a circular orbiter, assuming a ten-site global network.

Studies of direct orbit determination with down-looking GPS were carried out for NASA's geosynchronous Tracking and Data Relay Satellites.[37,43] With GPS orbit error assumed to be 7 m and clock error 2 m (SA off), a predicted three-dimensional TDRS position accuracy of 12 m is reached dynamically after 8 h, improving to 7–9 m after 24 h. With SA assumed on, the error jumps to over 100 m at 8 h, declining to 60–75 m at 24 h.

A ground tracking network of 12 GPS receivers can provide submeter GPS orbit accuracy. If those receivers were to track an additional high-altitude beacon,

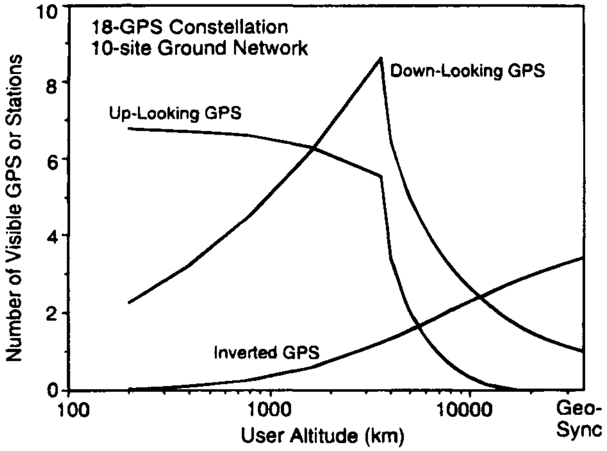


Fig. 10 Average numbers of GPS satellites visible with upward and downward looking hemispherical antennas and average number of sites from a 10-site ground network that can track an orbiting beacon, as a function of altitude.

all orbits could be estimated in one solution with comparable relative accuracy, scaled for distance. The same studies examined this inverted GPS tracking for TDRS. With assumed data errors of 25 cm for pseudorange and 1 cm for carrier phase at 30-min intervals, and a six-site ground network, the predicted three-dimensional rms orbit error over a 24-h dynamic solution arc was about 3 m for the TDRS satellites.

F. Highly Elliptical Orbiters

Because the preferred tracking modes and solution techniques differ for high and low orbiters, the application to highly elliptical orbiters, which may descend to a few hundred kilometers and rise to tens of thousands, presents a special challenge. Up- and down-looking GPS combined with ground-based Doppler during the high-altitude phase can provide particularly strong coverage. The proposed MUSES-B spacecraft, part of the Japanese very long baseline interferometry (VLBI) Space Observatory Program, was studied by Ref. 43. The MUSES-B would move from a perigee of 1000 km to an apogee of 20,000 km. The investigators applied a reduced dynamic strategy while the satellite was below 2000 km, and a purely dynamic strategy elsewhere. Combined omnidirectional differential GPS and ground-based Doppler gave a predicted orbit error of 50 cm for all position components at apogee, falling to less than 10 cm at perigee.

A similar mission, the proposed International VLBI Satellite, would have a perigee of 5000 km, enabling a purely dynamic solution around the orbit. Reference 43 found that with omnidirectional differential GPS and ground-based Doppler, as the apogee increases from 40,000 to 150,000 km, position error at apogee increases from about 15 cm to over 2 m. Such accuracy is needed for only select missions. Direct GPS orbit determination with omnidirectional reception could provide 10 m or better accuracy for nearly all highly elliptical orbiters.

VI. Dealing with Selective Availability and Antispoofing

The two GPS security features, antispoofing (AS) and selective availability (SA), can pose problems for SPS users. Over the years, various strategies have been devised to address them.

A. Antispoofing

Antispoofing is the encryption of the *P* code to prevent mimicking of the signal by others. In the presence of AS, a conventional SPS receiver would be able to track only the C/A code, recovering pseudorange and carrier phase on *L*1 only. This would prohibit computation of the standard dual-frequency ionospheric correction. The ionospheric effect, of course, depends on the altitude and field of view of the user. Above about 1000 km (assuming an upward-directed hemispherical field of view), the rare ionosphere permits submeter single-frequency orbit accuracy, even with no correction. At lower altitudes, as shown on EUE, the GRAPHIC calibration used in a reduced dynamic differential solution can give an orbit accuracy of about 1 m. That approach holds promise in both real-time and postprocessing uses for all but the most demanding requirements.

Low orbiters seeking subdecimeter performance must turn to dual-frequency calibration. There are several GPS receiver designs that operate in a codeless or quasicodeless mode; that is, they produce carrier phase and pseudorange at both frequencies without knowledge of the precise codes. Although codeless data are of degraded precision, the typical codeless phase error of about 1 cm in 1 s is consistent with the assumptions used in the studies presented here. Phase measurement noise, moreover, is generally not the dominant error in an orbit solution. Therefore, tracking performance will be largely unaffected by a switch to codeless operation.

B. Selective Availability

Selective availability consists of two measures to degrade positioning accuracy to the unauthorized user: the insertion of errors into the broadcast ephemeris and clock parameters, and "dithering" of the fundamental oscillator. Neither of these measures poses a problem for differential GPS tracking as we have defined it here. Receiver sampling times can be synchronized so that dither effects are common to all measurements and drop out of the differential solution. When sampling is not synchronized, quadratic interpolation to a common epoch can still achieve a high degree of dither cancellation, provided the sampling interval is no longer than about 30 s.[38] Ephemeris and clock errors do not come into play because those quantities are solved for with the ground data, either together with the user orbit or in advance.

Real-time direct GPS users encounter more difficulty. Those without access to accurate GPS orbits computed elsewhere will have to rely primarily on dynamic smoothing to reduce the effects of the ephemeris and clock errors and dithering. At nominal SA levels, the broadcast orbit error is 50 m or less on each satellite. (In recent years, the broadcast ephemeris has remained uncorrupted even when dither is active.) From the vantage of the user, the corrupted orbits will appear to some degree inconsistent with the GPS measurements and the user's dynamics—and the dynamic solution will then attenuate GPS orbit error. Early studies suggest that the error reduction will be about a factor of two. A 50-m GPS error may yield a 25-m user error. The actual reduction will depend on the solution strategy, receiver capacity, field of view, and other factors.

Smoothing of dither error by dynamic orbit estimation has been more thoroughly analyzed. Reference 39 simulated the dither process to examine the dynamic error attenuation as a function of the arc length. Figure 11 shows the net three-dimensional error (caused by dither only) for smoothing periods ranging from zero (point positioning) to 6 h. The receiver is assumed to track all satellites within a hemisphere, and dither is set at its nominal level. We see that with no smoothing, the rms dither error is about 30 m. After 2 h, this falls to 5 m, and after 6 h to less than 3 m. In a real-time application, in which the current state estimate cannot be smoothed with future data, the filtered dither error could be two or three times larger than shown in Fig. 11. Several investigators have attempted to mitigate dither effects by estimating a time-varying (stochastic) range bias to each GPS satellite using stochastic models tuned with actual SA data. But rapid dither variations are all but unobservable when the user state and clock are also estimated; such strategies tend, therefore, to sap data strength and degrade the orbit solution. Simple dynamic orbit estimation with a constant phase bias adjustment is generally more effective.[40]

A satellite such as TOPEX/Poseidon, which has well-modeled dynamics, can realize the full benefit of dynamic SA smoothing. At altitudes below about 600 km, dynamic model error will begin to offset the gain from smoothing, and at typical Shuttle altitudes the optimal direct solution may be little better than the point position solution under SA. To improve real-time accuracy at low altitudes, some form of near-real-time correction must be applied. This could be carried out, as is now commonly done for air and surface navigation, with

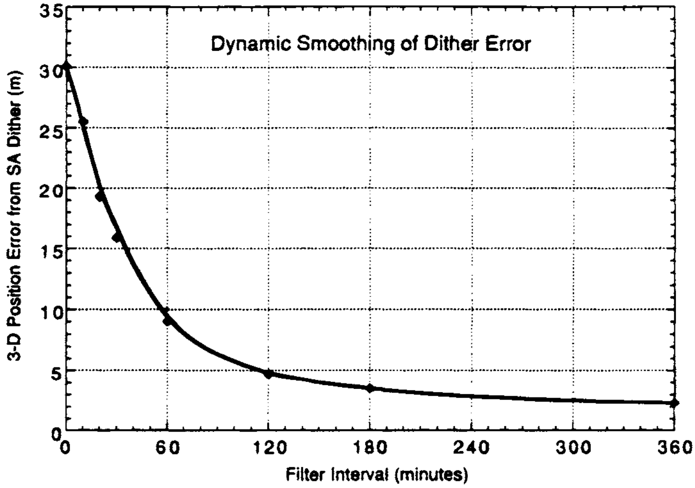


Fig. 11 Simulation results showing the three-dimensional position error that results from SA dithering, set at its nominal level, as a function of dynamic smoothing interval. No other errors are shown.

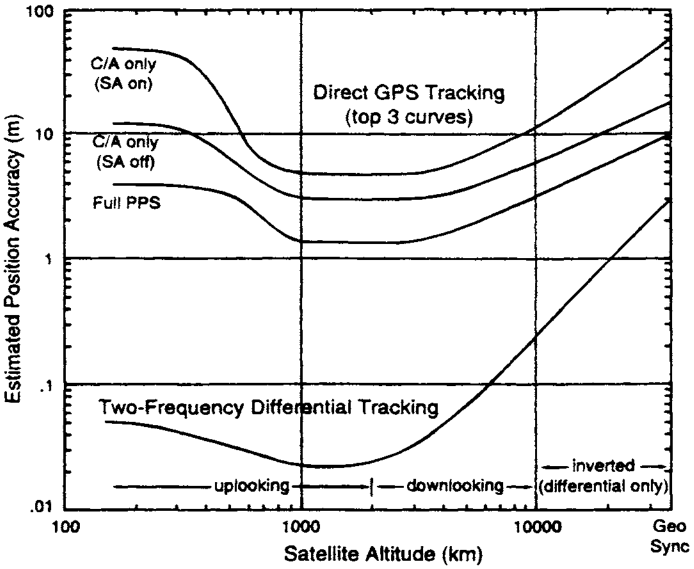


Fig. 12 Summary of the estimated orbit accuracies currently achievable with both differential and real-time direct GPS techniques. The direct solutions assume the use of precomputed GPS orbits of 2–3 m accuracy.

pseudorange corrections broadcast directly to users. Systems may soon be in place to send such corrections over wide areas through geosynchronous satellites (see Chapter 1 of this volume). A low orbiter equipped to receive those corrections could then achieve real-time position accuracy of about 1 m under SA.[29]

VII. Summary

The positioning strength provided by GPS is transforming Earth satellite orbit determination. With even the simplest receiving equipment it is now possible to determine the position of a low orbiter instantaneously to tens of meters, sufficient to meet the needs of most missions. The classical framework of dynamic orbit estimation can be adapted for GPS-equipped satellites in virtually any orbit to deliver orbit accuracies beyond the previous state of the art. Many low Earth satellites now in the planning stages will carry GPS for basic navigation and timing, and in some cases for direct scientific uses. Before GPS, orbits below 700 or 800 km could not be considered for satellites seeking accuracies of a decimeter or better. GPS promises to deliver few-centimeter accuracy at the lowest altitudes and for the most dynamically ill-behaved platforms. This creates the opportunity for low-power, low-mass, low-cost altimetry at an altitude of a few hundred kilometers, and for demonstrating precise sensing instruments on the Space Shuttle.

Figure 12 summarizes the performance that can be achieved as a function of altitude for both real time direct and after-the-fact differential GPS-based orbit determination. The curves reflect the optimal estimation strategy for each case.

For satellites above 10,000 km the standard differential technique is replaced by inverted GPS, where the orbiter carries a beacon tracked from the ground. The differential curve is consistent with the assumption of a dual-frequency codeless (SPS) receiver and is, therefore, unaltered by the presence of SA (eliminated by differencing) or AS. All curves for direct estimation assume the use of high-quality (2–3 m) GPS orbits and clocks distributed by civilian services, rather than the broadcast ephemeris. Thus, only dither error is included in the SA-on case. These curves are necessarily approximate; actual performance will depend on specifics of the GPS tracking configuration and satellite dynamics. But they offer a glimpse of the new standard GPS brings to orbit determination for missions of every description.

Acknowledgments

I am grateful to Catherine Thornton, Sien-Chong Wu, Willy Bertiger, Stephen Lichten, Ron Muellerschoen, Kenneth Gold, and Penina Axelrad for valuable discussions and other contributions. Portions of the work described here were carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.

References

[1]Axelrad, P., and Parkinson, B. W., "Closed Loop Navigation and Guidance for Gravity Probe B Orbit Insertion," *Navigation,* Vol. 36, 1989, pp. 45–61.

[2]Hesper, E. T., Ambrosius, B. A. C., Snijders, R. J., and Wakker, K. F., "Application of GPS for Hermes Rendezvous Navigation," *Spacecraft Guidance, Navigation, and Control Systems,* European Space Agency, 1992, pp. 359–368.

[3]Axelrad, P., and Kelley, J., "Near-Earth Orbit Determination and Rendezvous Navigation Using GPS," *Proceedings, IEEE PLANS* '86, IEEE Position Location and Navigation Symposium (Las Vegas, NV), Institute of Electrical and Electronic Engineers, New York, Nov. 4–7, 1986, pp. 184–191.

[4]Chao, C. C., Bernstein, H., Boyce, W. H., and Perkins, R. J., "Autonomous Stationkeeping of Geosynchronous Satellites Using a GPS Receiver," *Proceedings of the American Astronautical Society/AIAA Astrodynamics Conference* (Hilton Head, SC), AIAA, Washington, DC, Aug. 10–12, 1992, AIAA CP-92-4655, pp. 521–529.

[5]Lichten, S. M., Edwards, C. D., Young, L. E., Nandi, S., Dunn, C., and Haines, B. J., "A Demonstration of TDRS Orbit Determination Using Differential Tracking Observables from GPS Ground Receivers," *Proceedings of the Third American Astronautical Society/AIAA Spaceflight Mechanics Meeting* (Pasadena, CA), American Astronautical Society, San Diego, CA, Feb. 22–24, 1993, AAS Paper 93–160.

[6]Yunck, T. P., Wu, S. C., Wu, J. T., and Thornton, C. L., "Precise Tracking of Remote Sensing Satellites with the Global Positioning System," *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 28, 1990, pp. 108–116.

[7]Schreiner, W. S., Born, G. H., Larson, K. M., and MacDoran, P. F., "Error Analysis of Post-Processed Orbit Determination for the Geosat Follow-On Altimetric Satellite Using GPS Tracking," *Proceedings of the AIAA/American Astronautical Society Astrodynamics Conference* (Hilton Head, SC), AIAA, Washington, DC, Aug. 10–12, 1992, AIAA CP-92-4435, pp. 124–130.

[8]Munjal, P., Feess, W., and Ananda, M. P. V., "A Review of Spaceborne Applications of GPS," *Proceedings of the Institute of Navigation GPS '92,* Institute of Navigation, Washington, DC, 16–18 Sept., 1992, pp. 813–823.

[9]Parkinson B. W., "Navstar Global Positioning (GPS)," *Proceedings of the National Telecommunications Conference,* Nov. 1976.

[10]Farr, J. E., "Space Navigation Using the Navstar Global Positioning System," *Rocky Mountain Guidance and Control Conference* (Keystone, CO), American Astronautical Society, San Diego, CA, Feb. 24–28, 1979, AAS Paper 79–001.

[11]Van Leeuween A., Rosen, E., and Carrier, L. "The Global Positioning System and Its Applications in Spacecraft Navigation," *Navigation,* Vol. 26, 1979, pp. 204–221.

[12]Tapley, B. D., "A Study of Autonomous Satellite Navigation Methods Using the Global Positioning System," NASA-CR-162635, Dept. of Aerospace Engineering, Univ. of Texas at Austin, Austin TX, 1980.

[13]Wooden, W. H., and Teles, J., "The Landsat-D Global Positioning System Experiment," *Proceedings of the AIAA/American Astronautical Society Conference* (Danvers, MA), AIAA, New York, Aug. 11–13, 1980, AIAA CP-80-1678.

[14]Kurshals, P. S., and Fuchs, A. J., "Onboard Navigation—The Near-Earth Options," *Proceedings of the Rocky Mountain Guidance and Control Conference* (Keystone, CO), American Astronautical Society, San Diego, CA, Feb. 17–21, 1981, pp. 67–89.

[15]Masson, B. L., Ananda, M. P. V., and Young, J., "Functional Requirements of the next Generation Spaceborne Global Positioning System (GPS) Receivers," *Proceedings of the Institute of Electrical and Electronics Engineers National Telesystems Conference* (Galveston, TX), Institute of Electrical and Electronics Engineers, New York, 1982.

[16]Jorgensen, P., "Autonomous Navigation of Geosynchronous Satellites Using the Navstar Global Positioning System," *Proceedings of the National Telesystems Conference, NTC '82,* Galveston, TX, Institute of Electrical and Electronics Engineers, New York, Nov. 7–10, 1982, pp. D2.3.1–D2.3.6.

[17]Heuberger, J., and Church, L., "Landsat-4 Global Positioning System Navigation Results," *Proceedings of the American Astronautical Society/AIAA Astrodynamics Conference, Part I* (Lake Placid, NY), American Astronautical Society, San Diego, CA, Aug. 22–25, 1983 (AAS Paper 83–363), pp. 589–602.

[18]Fang, B. T., and Seifert, E., "An Evaluation of Global Positioning System Data for Landsat-4 Orbit Determination," *Proceedings of the AIAA Aerospace Sciences Meeting* (Reno, NV), AIAA, New York (AIAA CP-85-0286), Jan. 14–17, 1985.

[19]Ondrasik, V. J., and Wu, S. C. "A Simple and Economical Tracking System with Sub-Decimeter Earth Satellite and Ground Receiver Position Determination Capabilities," *Proceedings of the Third International Symposium on the Use of Artificial Satellites for Geodesy and Geodynamics,* Ermioni, Greece, Sept. 1982.

[20]Ananda, M. P., and Chernick, M. R., "High-Accuracy Orbit Determination of Near-Earth Satellites Using Global Positioning System (GPS)," *Proceedings, IEEE PLANS '82,* IEEE Position Location and Navigation Symposium (Atlantic City, NJ), Institute of Electrical and Electronics Engineers, New York, Dec. 6–9, 1982, pp. 92–98.

[21]Wu, S. C. "Orbit Determination of High-Altitude Earth Satellites: Differential GPS Approaches," *Proceedings of the First International Symposium on Precise Positioning with the Global Positioning System,* Rockville, MD, April, 1985.

[22]Yunck, T. P., Melbourne, W. G., and Thornton, C. L., "GPS-Based Satellite Tracking System for Precise Positioning," *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 23, July, 1985, pp. 450–457.

[23]Bierman, G. J., *Factorization Methods for Discrete Sequential Estimation,* Academic Press, Orlando, FL, 1977.

[24]Thornton, C. L., "Triangular Covariance Factorizations for Kalman Filtering," Jet Propulsion Laboratory Internal Document TM 33-798, Jet Propulsion Laboratory, Pasadena, CA.

[25]Lichten, S. M., and Bertiger, W. I., "Demonstration of Sub-Meter GPS Orbit Determination and 1.5 Parts in 10^8 Three-Dimensional Baseline Accuracy," *Bulletin Géodésique,* Vol. 63, 1989, pp. 167–189.

[26]Wu, S. C., Bertiger, W. I., Border, J. S., Lichten, S. M., Sunseri, R. F., Williams, B. G., Wolff, P. J., and Wu, J. T., "OASIS Mathematical Description," Vol. 1.0, Jet Propulsion Laboratory Internal Document D-3139, Jet Propulsion Laboratory, Pasadena, CA, 1986.

[27]Lear, W. M., "Range Bias Models for GPS Navigation Filters," Rept. JSC-25857, NASA Johnson Space Center, Houston, TX, 1993.

[28]Hatch, R. R., "The Synergism of GPS Code and Carrier Measurements," *Proceedings of the Third International Geodetic Symposium on Satellite Doppler Positioning,* Las Cruces, NM, 1982, pp. 1213–1231.

[29]Wu, S. C., and Yunck, T. P., "Precise Kinematic Positioning with Simultaneous GPS Pseudorange and Carrier Phase Measurements," *Proceedings, Institute of Navigation National Technical Meeting* (Anaheim, CA), Jan. 18–20, 1995.

[30]Wu, J. T., "Orbit Determination by Solving for Gravity Parameters with Multiple Arc Data," *Journal of Guidance, Control, and Dymanics,* Vol. 15, 1992, pp. 304–313.

[31]Bertiger, W. I., and Yunck, T. P., "The Limits of Direct Satellite Tracking with GPS," *Navigation,* Vol. 37, 1990, pp. 65–79.

[32]Wu, J. T., "Elimination of Clock Errors in a GPS-Based Tracking System," *Proceedings of the AIAA/American Astronautical Society Astrodynamics Conference* (Seattle, WA), AIAA, New York, 1984 (AIAA CP-84-2052).

[33]Yunck, T. P., Bertiger, W. I., Wu, S. C., Bar-Sever, Y., Christensen, E. J., Haines, B. J., Lichten, S. M., Muellerschoen, R. J., Vigue, Y., and Willis, P., "First Assessment of GPS-Based Reduced Dynamic Orbit Determination on Topex/Poseidon," *Geophysical Research Letters,* Vol. 21, 1993, pp. 541–544.

[34]Schutz, B. E., Tapley, B. D., Abusali, P. A. M., and Rim, H. J., "Dynamic Orbit Determination Using GPS Measurements from Topex/Poseidon," *Geophysics Research Letters,* Vol. 21, 1993, pp. 2179–2182.

[35]Rosborough, G., and Mitchell, S. "Geographically Correlated Orbit Error for the Topex Satellite Using GPS Tracking," *Proceedings of the AIAA/American Astronautical Society Astrodynamics Conference, Part 2* (Portland OR), AIAA, Washington, DC, Aug. 20–22, 1990 (AIAA CP-90-2956), pp. 655–663.

[36]Gold, K., Bertiger, W. I., Wu, S. C., Yunck, T. P., Muellerschoen, R. J., and Born, G. H., "GPS Orbit Determination for the Extreme Ultraviolet Explorer," *Proceedings of the Institute of Navigation GPS '93,* Salt Lake City, UT, 1993, pp. 257–268.

[37]Haines, B. J., Lichten, S. M., Malla, R. P., and Wu, S. C., "Application of GPS Tracking Techniques to Orbit Determination of TDRS," *Flight Mechanics/Estimation Theory Symposium,* Goddard Space Flight Center, Greenbelt, MD, 1992, (NASA CP-3186), pp. 117–128.

[38]Wu, S. C., Bertiger, W. I., and Wu, J. T., "Minimizing Selective Availability Error on Topex GPS Measurements," AIAA/American Astronautical Society Conference, Portland, OR, 1992, AIAA CP-90-2942.

[39] Bar-Sever, Y., Yunck, T. P., and Wu, S. C., "GPS Orbit Determination and Point Positioning Under Selective Availability," *Proceedings of the Institute of Navigation, GPS '90,* Colorado Springs, CO, Institute of Navigation, Washington, DC, 1990, pp. 255–265.

[40] Lear W. M., Montez, N. M., Rater, L. M., and Zyla, L. V., "The Effect of Selective Availability on Orbit Space Vehicles Equipped with SPS GPS Receivers," *Proceedings of the Institute of Navigation Conference, GPS '92,* Institute of Navigation, Washington, DC, 1992, pp. 825–840.

[41] Lichten, S. M., "Estimation and Filtering for High-Precision GPS Positioning Applications," *Manuscripta Geodaetica,* Vol. 15, 1990, pp. 159–176.

[42] Wu, S. C., Yunck, T. P., and Thornton, C. L., "Reduced-Dynamic Technique for Precise Orbit Determination of Low Earth Satellites," *Journal of Guidance, Control, and Dynamics,* Vol. 14, No. 1, 1991, pp. 24–30.

[43] Wu, S. C., Yunck, T. P., Lichten, S. M., Haines, B. J., and Malla, R. P., "GPS-Based Precise Tracking of Earth Satellites from Very Low to Geosynchronous Orbits, *Proceedings of the National Telesystems Conference* (Ashburn, VA), 1992.

[44] Yunck, T. P., "Coping with the Atmosphere and Ionosphere in Precise Satellite and Ground Positioning," *Environmental Effects on Spacecraft Positioning and Trajectories,* edited by A. V. Jones, Geophysical Monograph 73, IUGG Vol. 13, American Geophysical Union, Washington, DC, 1993, pp. 1–16.

[45] MacDoran, P. F., "A First-Principles Derivation of the Differenced Range Versus Integrated Doppler (DRVID) Charged Particle Calibration Method," *JPL Space Programs Summary 37-62,* Vol. II, March 1970.

Chapter 22

Test Range Instrumentation

Darwin G. Abby*
Intermetrics, Inc., Holloman Air Force Base, New Mexico 88330

I. Background

HISTORICALLY, land-based test facilities have used a combination of radar, distance-measuring equipment (DME), optical trackers such as cinetheodolites, and other miscellaneous instrumentation to provide time–space position information (TSPI) to satisfy test platform positioning requirements. In the early 1970s, laser trackers became available to support test activities. Each of these systems had their strong and weak points, and systems were used depending on the accuracy, area of coverage requirements, and cost considerations. Radar systems could cover fairly large line-of-sight areas, but accuracys were low (25–50 ft), and cost was high. DME and laser systems had limited areas of coverage, were fairly accurate (2–5 m), and had medium cost. Cinetheodolites provided very high accuracy (0.5–1 m) over very limited areas; the cost was very high, and delays of 2–6 weeks for data processing was the norm.

To obtain overland coverage of larger areas, instrumented test ranges were selected at both ends of a flight trajectory and special radar sites were built or FAA air traffic radars were used to cover the enroute areas. Test support for these types of overland flights was very expensive. Examples of these types of support activities were the Edwards AFB, California to Utah Test Range located near Salt Lake City, Utah and a route from Edwards Air Force Base to White Sands Missile Range in New Mexico.

A combination of radar and DME [i.e., the General Dynamics, Inc. Position Location System (PLS) and the Cubic Corporation Air Combat Maneuvering Instrumentation (ACMI)] systems were used to satisfy combat training requirements. Most of these systems were designed to work in real time, and locations of ground transponders and communication links limited the area of coverage. These systems were very expensive to operate and maintain.

*Senior Staff Consultant, Navigation Systems Department, Systems and Software Applications Division.

The broad ocean test ranges had a much more difficult problem for obvious reasons. Radars or other types of tracking instrumentation could be located only on coasts, islands, or instrumented ships and aircraft. Highly instrumented terminal areas for missile testing were built using radar and sonobuoys for positioning impacts. These systems met the requirements at the time, but again costs to maintain these systems were very high.

The use of GPS for test and training applications began to surface in studies conducted in the 1979–1981 time frame. A system that could provide highly accurate TSPI over unlimited areas at low cost had the potential of solving a wide range of test and training requirements. In many cases, there were large geographic areas and even global test and training requirements that either could not be met by other candidate systems, or could only be met at very high cost when compared to the cost of a GPS-based system. At the time these studies were conducted, differential GPS was in its infancy and was not as highly developed as it is today. Therefore, GPS accuracies of 10–15 m Spherical Error Probable (SEP) were assumed for evaluating GPS utility for test and training.

In 1979, SRI International[1] completed a study that concluded that exploitation of GPS in the test community for test article TSPI measurement offered significant economic and operating advantages for test and evaluation (T&E) operations. The 10–15 m accuracy met many of the T&E requirements but did not meet positioning requirements for precision weapon system testing. The main advantage was cost reduction and increased flexibility. The potential to achieve the accuracys any where in the world by installing a GPS receiver brought a cost-effective capability to the test community.

Army Captain William Reinhart completed a thesis[2] in 1981 that investigated the use of GPS for replacing positioning systems used in training systems at Fort Ord, California. Operational battlefield training systems use instrumentation to determine who is firing at whom, simulate each engagement mathematically, and provide opposing players with realistic engagement results. A very accurate positioning system is required to support real-time casualty assessment calculations and data for evaluating the success of opposing armies. Tests were conducted, and 10–15 m rms GPS accuracy were demonstrated. Horizontal accuracy requirements were not met; however, the vertical accuracy was better. Advantages of GPS over existing systems were: 1) larger player capacity; 2) ability to perform concurrent experiments; and 3) exportability or ability for worldwide use.

A 1980 MITRE Corporation report[3] evaluated the use of Navstar GPS receiver equipment as Navy range instrumentation for R&D test and air and ground combat training. Five GPS receiver configurations were considered for satisfying test and training requirements ranging from high-velocity–high-acceleration systems, such as air-to-air missiles, to low-dynamic applications, such as soldiers and trucks. They concluded that the five GPS equipment configurations defined could provide the required accuracy over the range of dynamic conditions expected in test and training environments.

The overall conclusion from these studies was that GPS offered cost-effective solutions to most test and training (T&T) applications. With the advent of differential GPS (DGPS), even many "positioning requirements for precision weapon system testing"[3] could be achieved, and the horizontal requirements for training applications could be met. The use of GPS for T&T applications was only a matter of developing and supplying the systems.

The Undersecretary of Defense for Research and Engineering (USDR&E) established a tri-service steering committee (TSSC) in 1981 to evaluate the possible application of GPS as a TSPI source for test and training ranges.[4] The TSSC was also to evaluate test and training requirements and issues and to recommend interim and long-term test range applications of GPS. The TSSC concluded that GPS could satisfy test and training requirements and could significantly improve position and velocity measurement accuracy in most applications.

The TSSC recommended and the USDR&E approved the establishment of a joint program to develop a family of GPS range equipment specifically tailored to the needs of the test and training communities. The Air Force was chosen as executive service and, via AFSC, delegated to the Range Directorate (AD/YI), Missile Systems Division, Eglin Air Force Base, Florida, program management responsibility. In July 1983, AD/YI established the Range Applications Joint Program Office (RAJPO), to continue to analyze technical issues, manage a preliminary test program, evaluate range integration issues, and develop triservice equipment using GPS as the TSPI source for DOD test and training ranges. The RAJPO equipment is described later in the chapter.

A significant conclusion of the TSSC study was that GPS translators would be useful in a variety of range applications, particularly those involving destructive testing (e.g., missiles) and when the number of test articles active simultaneously is limited.[5] GPS translator technology was being developed at the time by the Navy Fleet Ballistic Missile agency for support of Trident I testing. Translators are now used extensively by the strategic missile community, and the technology has advanced significantly.

II. Requirements

TSPI requirements for test and training applications are very similar, and positioning systems developed for one or the other can potentially be used by both. The differences in the applications are primarily in data management and utilization.

A. Test Requirements

Testing of new weapon systems requires very accurate reference systems for position and velocity. In addition, requirements exist for acceleration and attitude truth references on many weapon systems. Department of Defense test ranges use a wide variety of systems including tracking radars, laser trackers, cinetheodolites, and range measurement systems to provide TSPI. Systems under test include such weapons as bombs and missiles, targeting systems, and navigation systems. In the past, most testing has been conducted on instrumented test ranges within a controlled area. Modern weapon systems using new technologies have created new and unique support requirements. The need for cost-effective, disposable TSPI instrumentation for one-time missile shots is one example. Missiles can vary from the small tactical high-speed aircraft missiles to such larger strategic ones as the Minuteman. Many modern systems require testing over extended areas of land and sea, where it can be very difficult to impossible to provide tracking instrumentation. An example would be evaluating active target sensing and recognition sensors in different terrain and climates. Most test programs do not require the high-accuracy TSPI in real time at the test vehicle or at a range control facility; however, in almost all cases, continuous real-time information on the test platform position must be available to satisfy safety requirements. Real-time requirements for position and/or velocity create significant technological and cost impacts.

Reference accuracy requirements vary depending on the weapon system. The rule of thumb is to have a truth reference of approximately a factor of ten more accurate than the system under test. However, the accuracy often is limited by the availability of a cost-effective system.

The TSSC, as a part of its charter, also conducted a survey of twenty-two DOD test ranges. One objective of this survey was to summarize the test and training accuracy requirements using six different parameters and to determine if GPS could meet those requirements. The results of the survey and what GPS can provide are shown in Table 1. At that time, GPS could meet most of the requirements except in the scoring area, where accuracies of 1–10 ft are required. DGPS accuracies of 0.1 m are now available.

The use of GPS for missile tracking presents a unique set of requirements. First, missiles are not recoverable, and the equipment is destroyed on every test which makes cost a major factor. In addition, size and weight are critical for missile applications, which has led to another GPS translator technology area. The two primary requirements for GPS missile-tracking support are for range safety and precise trajectory determination. For range safety, the missile position must be determined and made available in real time so that corrective action can be taken immediately in case of emergency. In general, precise trajectory determination can be performed postflight unless real-time scoring is required.

B. Training Requirements

Training requirements include instrumenting battlefield or air combat participants, (soldiers, tanks, trucks, helicopters, aircraft, etc.) in such a way that mock battles can be conducted. The question of who shot whom can be sorted out for real-time casualty assessment and made available to commanders and troops in real-time and/or postmission for analysis and debriefing. New tactics are evaluated during training exercises to determine effectiveness against new weapon systems. The major instrumentation components are a position- and velocity-determining system, data links and/or recorders, central processors, and displays. Position and velocity accuracy and data rate requirements are a function of participant dynamics (soldier vs F-16) and training objectives (tactics vs troop training).

Frequently, accuracy requirements are driven by the availability of cost-effective instrumentation. Developments continually are in process to try to improve accuracy and minimize cost. The U.S. Army Training and Doctrine (TRADOC), test and experimentation command (TEXCOM), formerly combat development experimentation command (CDEC), Fort Ord, California has been developing instrumentation for training applications for many years.

Table 1 TSPI requirements summary (22 ranges)

TSPI performance parameters	Training and OT&E[a] ranges				DT&E[b] and OT&E ranges			
	Air	Land	Sea (fixed)	Sea (moving)	Long-range	Extended range	Short-range (land)	Short range (water)
Real-time accuracy								
Position (ft)	⑤–200	15–30	200	1000	20–100	⑤–20	⑤–100	⑤–100
Velocity (fps)	0.1–15	3–9	100	——	0.5–5	5	1–20	1–20
Data rate (#/s)	1–20	1–10	5	1–5	20	20	1–100	1–20
Post-test accuracy								
Position (ft)	$\boxed{1}$–200	6–30	10	TBD	10–20	30	$\boxed{2}$–15	$\boxed{2}$–15
Velocity (fps)	0.1–15	3–9	0.1–5	——	0.01–0.1	0.01–0.02	0.1–10	0.1–10
Scoring (ft)	$\boxed{1–10}$	$\boxed{1–6}$	——	——	50	$\boxed{1}$	$\boxed{1–5}$	$\boxed{1–5}$
Number of test articles	1–90	2–2000	50	60–125	3–10	1	12–20	12–20
Coverage								
Altitude (kft)	0.1–100	0–10	0.1–58	0–60	300	0.1–30	0–100	0–100
Distance (nm)	30–60	30 × 30	75 × 75	350 × 500	150–3000	100 × 600	50 × 150	125 × 200

○ GPS marginal. □ GPS can not satisfy.

TEXCOM's most valuable contributions in weapon system analysis have historically been to detect strengths or weaknesses that were not intuitively obvious. Often these system traits become apparent only when the system was fielded in a combined arms environment, during which the soldiers and/or the weapons systems are task loaded and carry the threat of being killed. This requirement results in a need for high-accuracy, real-time position data. The most important requirement for TEXCOM's development efforts is to provide 1-m position accuracy for non–line-of-sight weapon systems that simulate conditions actually existing in the field. The immediate weapon candidate is the fiber optics guided missile (FOG-M).

III. Range Instrumentation Components

The three major components for a DGPS test range system are a GPS reference station, a data link, and test vehicle instrumentation. The functions of these systems are described in the sections on DGPS. The intent here is to address unique characteristics relative to test applications. The components and function of a translator system are also described.

A. GPS Reference Station

Most DOD test ranges have access to selective availability/antispoof (SA/AS) authorization, and as a result, they can operate in the GPS precise positioning service (PPS) mode. However, this requires that security procedures be implemented to protect classified data.

GPS standard positioning service (SPS) reference stations are also being used by test ranges because they are available off-the-shelf at low cost, and they eliminate the need for security procedures. Differential GPS accuracies of 5–8 m have been achieved using these systems, thus meeting many test requirements.[6]

The optimum GPS reference receiver should track all visible satellites simultaneously and provide corrections for any potential satellite the test platform receiver might track. In the case of a high dynamic aircraft, the tracked satellite subset could be almost any combination of those visible. The generic components of a DGPS reference station include the reference receiver with antenna, data processor and controller for real-time or postprocessing, data-recording system, keyboard and display, printer/plotter, and optionally an external atomic frequency standard. The reference station interfaces with a data link system for real-time operations. The data broadcast on the link are also recorded for any postmission processing requirements. The operator interface is required for initializing the system, controlling the reference receiver, and monitoring data quality. An optional meteorological data interface may be included.

The Radio Technical Commission Maritime (RTCM) has been the leader in attempting to standardize the format for differential corrections produced by a GPS reference station. The RTCM established Special Committee-104 to prepare a standardized format resulting in RTCM SC-104 of January 1, 1990.[7] Most commercial DGPS vendors have adopted the standard and use it for differential GPS systems. On the other hand, NATO, RAJPO and the Range Commanders Council are developing other standards for exchange of GPS data.

B. Data Links

Detailed descriptions of the different data links being used by the ranges are discussed later. In general, the commercial vendors are adapting off-the-shelf communications radios for data links. To accommodate multiple players they use a combination of time division multiple access techniques and multiple frequencies. RS-422 or RS-232 data ports on the receiver are connected to modems and then to the communication radio. The same interface to the remote user receiver is used in reverse. In almost all cases two-way communication is available.

The RAJPO is developing a complete customized system that interfaces with the suite of RAJPO equipment. Details are outlined in Sec. V.A.1.

Other potential links include cellular telephones, L- and S-band telemetry systems, existing weapon system data links, satellite communications, etc. Virtually any rf link could be used to transmit the differential GPS data. Data rates are very low, on the order of 50 baud, but the data link implementation and vehicle dynamics affect the final rate.

C. Test Vehicle Instrumentation

The GPS receiver and associated instrumentation on the test platform can vary significantly depending on the vehicle dynamics, accuracy requirements, data-processing mode (real-time or postmission), and use of PPS or SPS operations. The minimum package for a low dynamic platform would be the GPS receiver and a data recorder for postprocessing and TSPI generation. For a high dynamic platform, the GPS receiver would need to be aided with an inertial navigation system (INS). A receiver dataport would interface with either a recorder or a real-time data link.

D. Translator Systems

Historically, GPS single- or multi-channel receivers have been used on test vehicles. These receivers receive and process the GPS signals and output position and velocity estimates for use onboard the test vehicle or downlinked to additional equipment at the ground facility for display and data processing. Translators should be considered when the test mission has a requirement for time–space–position information (TSPI) at a ground facility and one of the following is true:

1) The program is concerned with the high cost of GPS receivers because the test vehicles will be attrited.

2) There is insufficient volume to support an onboard GPS receiver.

3) Detailed postmission analysis of the GPS data is required.

Various programs using GPS translators have been supported including the Strategic Defense Initiative Office (SDIO) ground-based interceptor (GBI) and the Peacekeeper.

Figure 1 shows the system concept for the GPS translator equipment. The system consists of a user-equipped translator and a ground-based translator processor system (TPS), of which the GPS translator receiver is an element.

The GPS L-band signals are received at the missile, translated to S-band, and then retransmitted to the ground station.[5] The composite of all GPS satellites

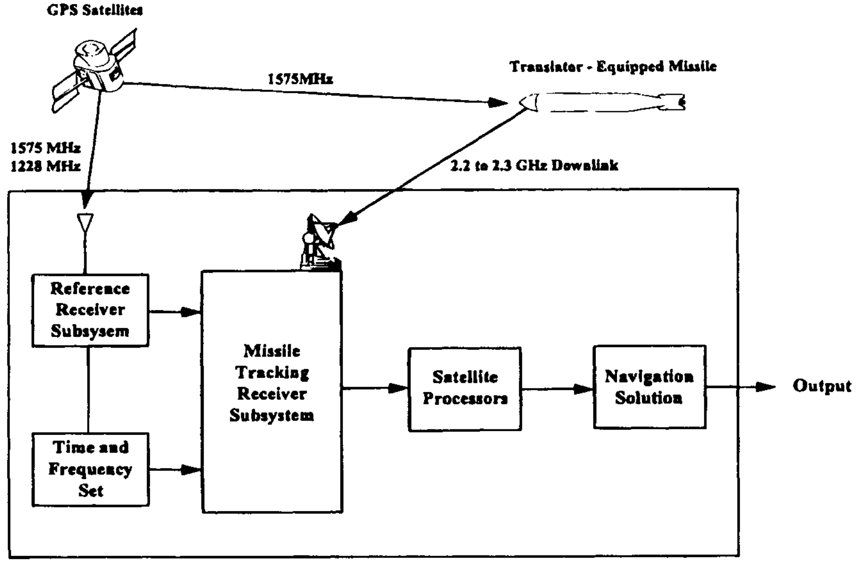


Fig. 1 Real-time missile tracking with GPS.

visible to the missile are received at the ground station for processing. A GPS reference receiver, collocated at the ground station, provides ground station location in GPS coordinates, GPS time, and satellite ephemerides. Because both an S-band target receiver and an L-band reference receiver are used at the ground station, the system becomes a differential navigation system with associated accuracy advantages. The system only uses the satellite L-1 link and C/A code through the missile translator. This configuration was selected for the following reasons:

1) The C/A code has 3 dB more power than the P code, providing a higher quality Doppler (velocity) measurement.

2) The C/A code provides sufficient position accuracy for most missile applications.

3) Translator power output required for the C/A code would have to be increased by a factor of 10 to use P code to retain the same downlink margin.

4) Use of the L-2 link through the translator would add a receiving antenna to the missile, a second receiver channel to the translator, and the required translator output would double.

The composite GPS satellite signals are filtered, amplified, and translated to one of several possible output frequencies in the 2200–2300 MHz range. This selectable output frequency feature allows several missiles to be tracked simultaneously by using frequency division multiple access (FDMA) on the missile to ground links. In addition, a pilot carrier is synthesized in the translator and introduced in close frequency proximity to the translated L-1 spectrum. The combined pilot carrier/translated L-1 is then transmitted to the ground station.

The signal from the missile is received at the ground station by a telemetry receiving antenna. This antenna is typically a high-gain parabolic reflector antenna, which is necessary to maintain a viable telemetry link at long ranges using minimal onboard telemetry transmitter power.

There are both advantages and disadvantages to using GPS translators. Translators are less expensive than equivalent GPS receiver (i.e., space-qualified, etc.). Translators also consume less power, occupy less space and—with a properly designed TPS—can acquire signals and provide a TSPI solution faster than GPS receivers. Translators can (again with a properly designed TPS) track GPS signals to a lower signal level than onboard GPS receivers (because of reference receiver aiding). In addition, a TPS is configured to support postmission processing, thus allowing GPS signals to be tracked through high accelerations (>20 *g*s), a real challenge for most GPS receivers. The key disadvantages are the large bandwidth downlink requirements (~2 MHz, C/A code; ~20 MHz, P code) and the need for high-gain telemetry antennas to track the user vehicle. A specific system being developed by the RAJPO is described in Sec. V.A.1.e.

E. Digital Translators

In the past four years, NAVSYS Corp. has taken advantage of digital technology to develop a smaller and less expensive system originally called the *Advanced Translator Processing System* and more recently called the *Digital Translator Processing System.*[8–10] The flight unit is 30 in.^{3}, weighs 3 lb, and uses only 28 W of power, vs the older analog units that range from 180 in.^{3} to the more recent analog design that is 40 in.^{3}; weighs 5 lb and requires 56 W of power.

Digital translators take advantage of recent developments in digital microwave radios. A block diagram of the vehicle translator and ATPS system are shown in Fig. 2. As in a conventional analog translator, the L-band GPS signal first passes through an L-band preamplifier. The signal is then filtered to select either the P-code or the C/A code bandwidth. The filtered L-band signal is next sampled and quantized by an A/D converter. The A/D outputs are used to modulate an S-band carrier. A pilot carrier (PC) is then added to the modulated signal. The combined signal is the digital translator output. The digital translator architecture eliminates the need for an IF frequency an significantly reduces the filtering requirements which results in significant savings in size, weight and power dissipation.

A key element of the advanced translator processing systems (ATPS) is the preamplifier/downconverter (P/DC) module developed by NAVSYS to condition the received translator signal so that it can be tracked by a conventional C/A code receiver. The P/DC tracks the pilot carrier from the GPS translator and uses this signal to down-convert the received S-band signal back to the L-band signal received at the translator. This signal can then be processed using a conventional off-the-shelf digital GPS receiver. There are very key functional similarities between the digital translator systems and the newer digital GPS receivers, as shown in Fig. 3.

There are other significant advantages of digital translators. The digital translator approach directly facilitates data encryption. The sampled and quantized L-band signal is a digital stream that can be encrypted by any conventional data encryption technique prior to modulation on the S-band carrier. In addition, the pilot carrier can be optionally modulated with telemetry data that could include

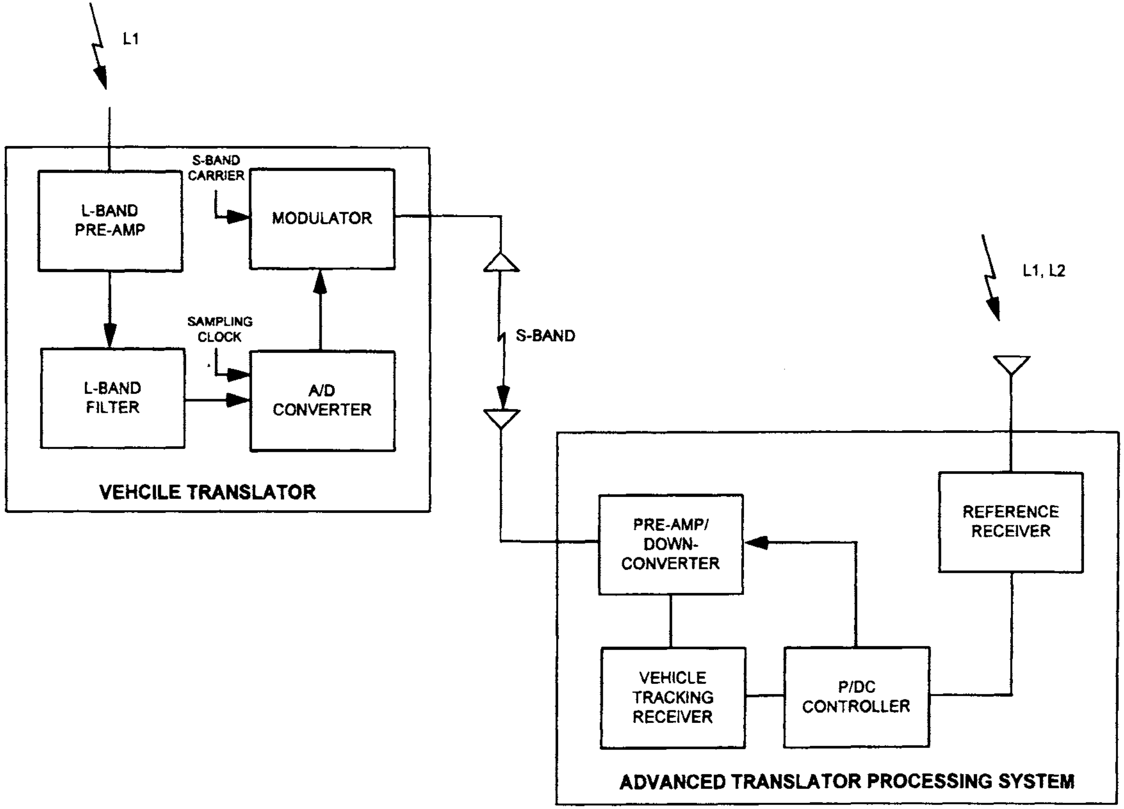


Fig. 2 Digital vehicle translator and translator processing system.

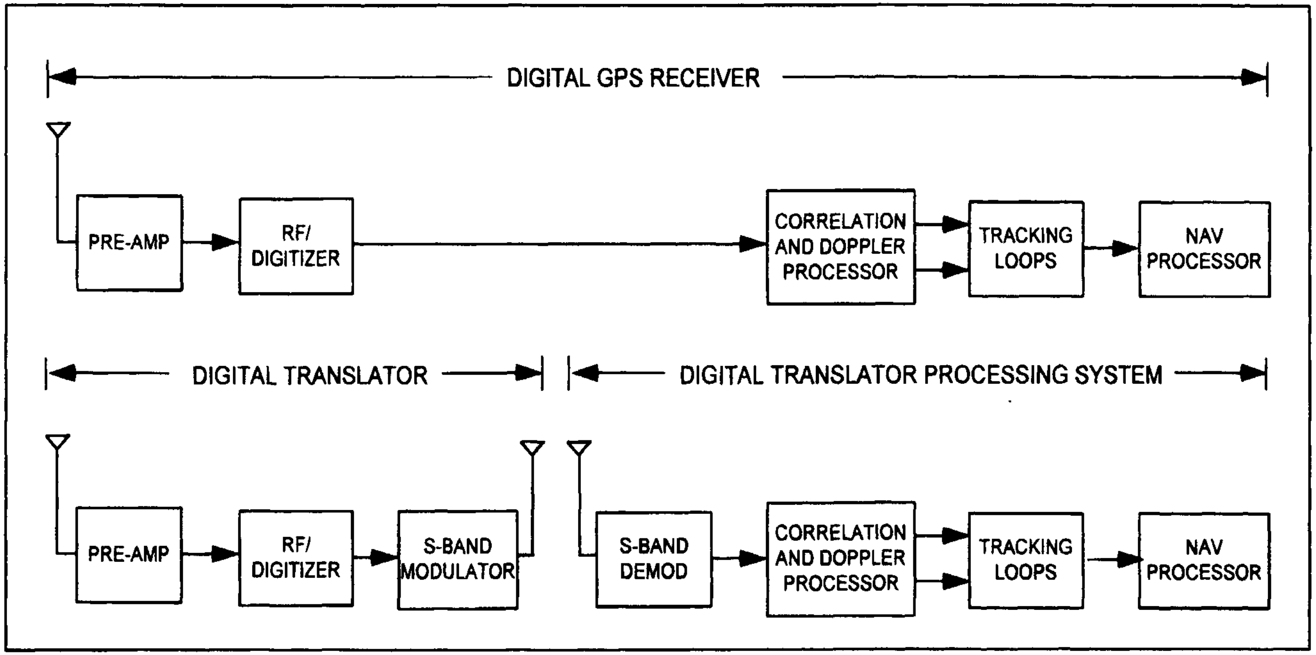


Fig. 3 Digital GPS reciever/digital translator comparison.

inertial reference unit (IRU) data. Intertial reference unit data can be used in post-processing navigation filters to aid the navigation during times of GPS outages.

IV. Differential Global Positioning Systems Implementations

The primary DGPS implementation method used by test range activities is one wherein raw data from the GPS receiver in the test vehicle are recorded or downlinked to the reference station, as shown in Fig. 4. The reference station computes the pseudorange corrections and rate of change of the corrections, applies the corrections to the test vehicle receiver's measurements, and computes the TSPI. This method takes advantage of a ground computer's larger capacity to perform more sophisticated processing and potentially achieve better accuracy. This method also offers the option of sending all data to the range host computer where the data can be combined with other range sensors, processed in a large Kalman filter, and even smoothed to achieve improved results. The data to be downlinked include: pseudoranges and pseudorange rates, satellite ID numbers age of data, ephemeris (AODE), receiver identification, user time, measurement quality estimates (if available) and INS data (attitude, velocities and/or accelerations). In this mode, the link is only one way and the test vehicle does not have the final solution, which in most test programs is not required. If TSPI is not required in real time, l the data can be recorded and postmission processed. This option 1) eliminates the need for a data link, 2) simplifies the instrumentation significantly, 3) eliminates the line-of-sight restriction required for RF links, and 4) allows the test vehicle to cover much larger geographic areas to satisfy many modern weapon system requirements.

V. Existing Systems

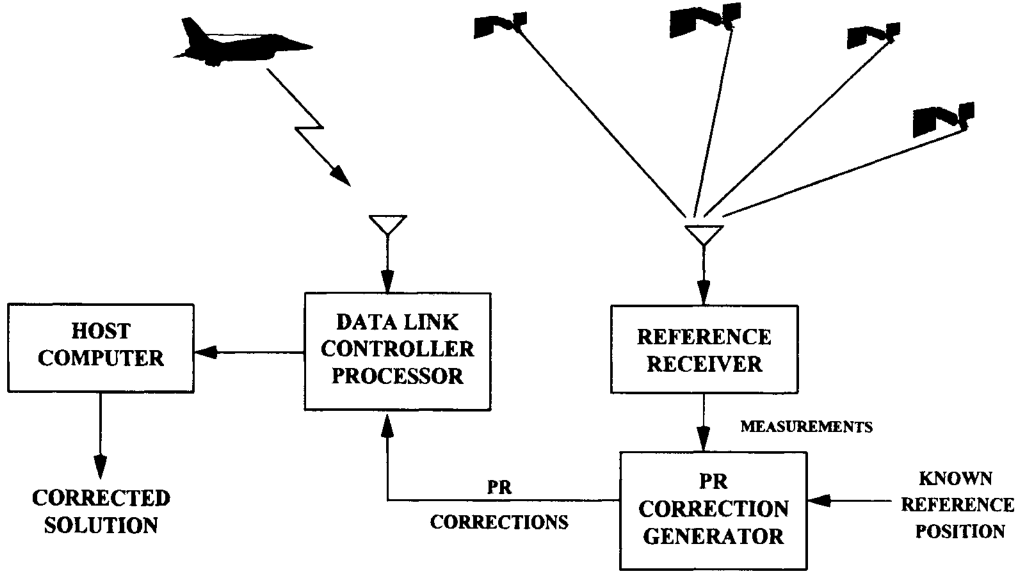
Development of DGPS and the applications for test ranges has been ongoing for over 9 years. There are several systems of both PPS and SPS that currently are operational or undergoing acceptance testing. These systems are described in this section.

A. Department of Defense Systems

Department of Defense test facilities have access to the PPS. The GPS receivers used are keyable and remove the effects of SA/AS. The two U.S. Air Force agencies that develop this type of equipment are the RAJPO and the GPS Joint Program Office.

1. Global Positiong Systems Range Applications Joint Program Office

The RAJPO is developing a family of equipment to use the Navstar GPS on DOD test and training ranges to provide TSPI. The individual components are designed so that range systems can be assembled to meet a variety of requirements. The family of equipment includes the high dynamic instrumentation set (HDIS), a reference receiver/processor (RR/P), C/A code receiver (CACR) and a data link system (DLS). Other equipment being developed for range applications



- VEHICLE GPS MEASUREMENTS DOWNLINKED & SENT TO HOST
- THE PRCG GENERATES CORRECTIONS FOR ALL SVs IN VIEW & SEND TO HOST
- HOST PROCESSES DATA AND GENERATES CORRECTED SOLUTION

Fig. 4 Differential GPS method 3.

include translators and translator processing systems and ground transmitters (GTs). A brief description of each component follows.

a. High dynamic instrumentation set (HDIS). The HDIS is five-channel, fully authorized for PPS, GPS receiver designed to fit in a 5-in. air incercept missile (AIM)-9 pod. It includes the antenna system, receiver/processor, range flexible modular interface (RFMI), power conditioning, and all interconnecting cables and connectors. The HDIS has provisions for integration with an IRU, a control display unit (CDU) for operator interface, and support equipment for maintenance. In addition to being able to receive and process GPS satellite navigation signals, the unit can also receive and process GT signals. The HDIS is capable of operating in a differential GPS mode by accepting differential corrections linked from the RR/P.

The integration of an HDIS, an IRU, data link translator (DLT) and solid-state recorder into a 5-in. AIM-9 missile pod provides a small autonomous instrumentation package that can be mounted on aircraft wing weapon stations. The same set of equipment can be configured for a small pallet for installation inside an aircraft. The combination of a pod or pallet system, RR/P and a DLS is called the advanced range data system (ARDS).

b. Reference receiver/processor. The RR/P includes an HDIS receiver, a pseudorange correction generator (PRCG), a navigation correction processor (NCP), an RFMI, a meteorological sensor subsystem, and a control display subsystem. The RR/P provides data outputs to the DLS for transmission to the test or training vehicles and to a host range computer system for processing. The RR/P tracks up to eight satellites and provides pseudorange corrections, rate of change of pseudorange corrections, raw pseudorange, meterological and satellite message data on all satellites. The RR/P operates in either the authorized or unauthorized mode.

c. C/A code receiver. The CACR is a small commercial C/A code receiver that has been modified to interface with all the RAJPO range systems. It will be used primarily for ground vehicle and manpack test and training applications.

d. Range data link (RDL). The RDL operates in segments of the 1350–1530 MHz (L-band) telemetry band; each RDL net uses two frequency channels within that band as shown in Fig. 5. Within a given area, multiple nets can be operated if sufficient channel pairs are available. Multiple nets can be operated independently, if applications are separate and disjoint, or their operation can be coordinated to expand the capacity of the system in a single large application.

An individual RDL net can contain up to 2000 RDL transceivers, ground and airborne. Operation is segmented in time into 330 time slots per second, during each of which a single transceiver transmits one message to one or more of the other transceivers. A transceiver can be assigned as many time slots as necessary to fulfill its communications needs. The assignment of transceivers to time slots is performed automatically by the system in accordance with needs, and it changes dynamically as those needs change. Needs can include relaying if the originator of the message is not within communications range of the recipient, and a small fraction of total net capacity (generally 10% or less) is taken up by internal messages used to control such assignment and reconfiguration.

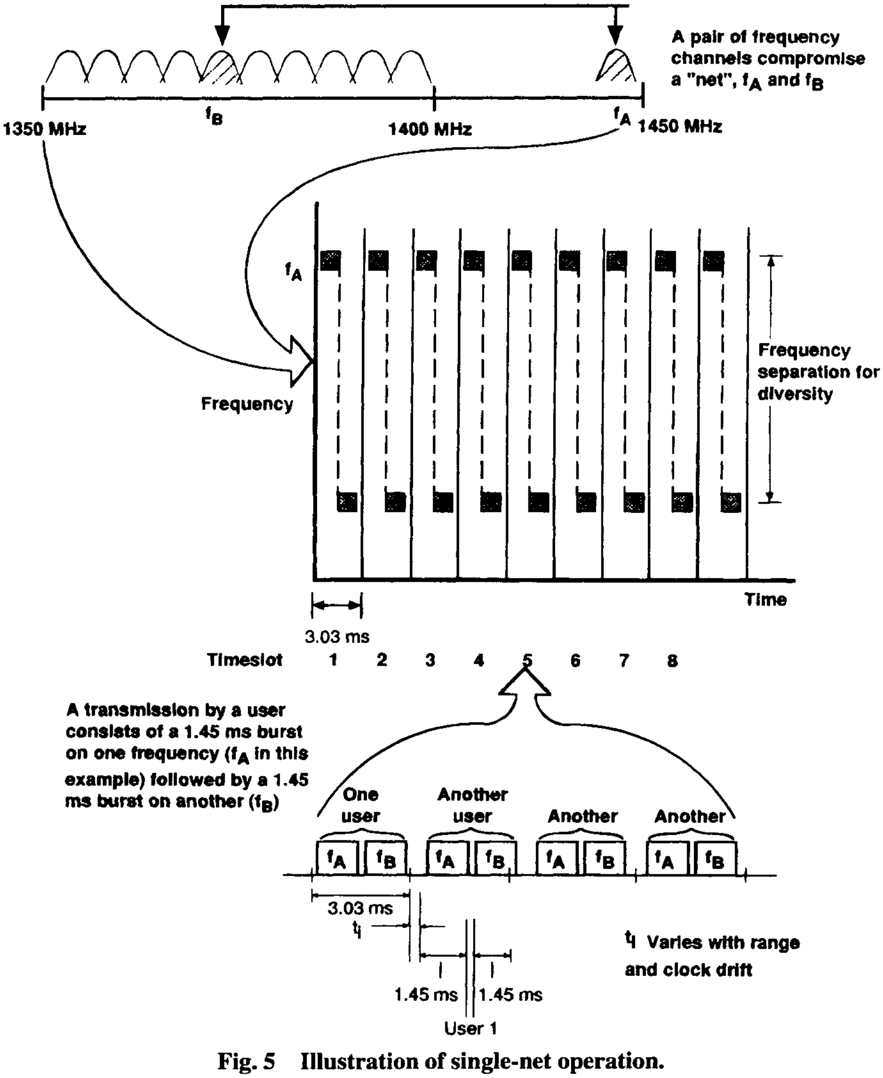

A pair of frequency channels compromise a "net", f_A and f_B
1350 MHz
f_B
1400 MHz
f_A
1450 MHz
f_A
Frequency
Frequency separation for diversity
Time
3.03 ms
Timeslot 1 2 3 4 5 6 7 8
A transmission by a user consists of a 1.45 ms burst on one frequency (f_A in this example) followed by a 1.45 ms burst on another (f_B)
One user
Another user
Another
Another
f_A f_B
3.03 ms
t_i
1.45 ms
1.45 ms
User 1
t_i Varies with range and clock drift

Fig. 5 Illustration of single-net operation.

In multinet coordinated operation, specific common time slots are assigned to the control function on a common frequency channel pair. During the remainder of the time slots, data transmissions from different participants would occur simultaneously on the various assigned channel pairs.

The configuration of the system can be tailored to individual ranges and further to particular test scenarios within those ranges. For example, a small-scale developmental test of a new vehicle with custom telemetry-gathering systems on board that produce, say, 28 kb/s, could be accommodated by interfacing those systems to the data link, defining 700-bit messages to carry the data, and assigning 40 slots per second to this function. An additional 10 slots per second might be

40 slots per second to this function. An additional 10 slots per second might be assigned to downlinking TSPI data from the platform, and three other platforms might also be tracked, at a 10 slot per second rate. Uplink control might require an additional 10 slots per second for the test platform, and for each platform. System control might require a total of 20 slots per second. These figures total 140 slots per second, hence, the system is operating at 42% of the full capacity of one net. Some of that excess capacity might be used for relaying; if the test vehicle were not within line-of-sight of any ground station, an additional 60 slots would be required for that function.

Hardware—Data link units are designed for mounting in an AIM-9 (5-in. diameter) pod as shown in Fig. 6. The 20-W transceiver occupies 14.5 in. An optional 60-W high-power transceiver occupies 18 in. of rail space within the pod, and the associated power supply, an additional 12 in. Antennas for pod-mount and internal aircraft installation are nominally omnidirectional. Stacked dipole antennas are planned for ground applications, omnidirectional in azimuth, but with elevation-plane beamwidths of about 13 deg. Most links within the system will connect airborne units and ground stations.

System control—Up to 17 ground stations are connected to a central facility, called the Data Link Controller (DLC). A variety of connecting links may be employed, using existing range communications systems such as microwave and wireline. The DLC and ground stations perform an accounting/error detection/retransmission protocol in their communication over these connecting links to protect against errors generated in them.

e. GPS translator equipment. The Interstate Electonics Corporation under contract to the RAJPO is one of the major developers of translator equipment. Fig. 7 shows the block diagram for the RAJPO analog translator.

Translator—1) *Operation*: Global positioning system satellites output two primary frequencies, denoted L_1 (1575 MHz) and L_2 (1228 MHz). The L_1 frequency contains two orthogonal pseudorandom codes, the (C/A) code and the precise (P) code. The L_2 frequency contains only the P code. The translator front-end

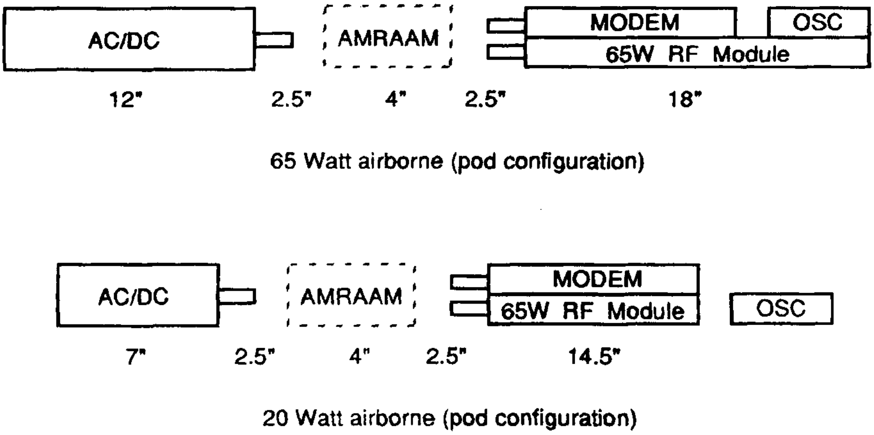


Fig. 6 Data link pod configuration.

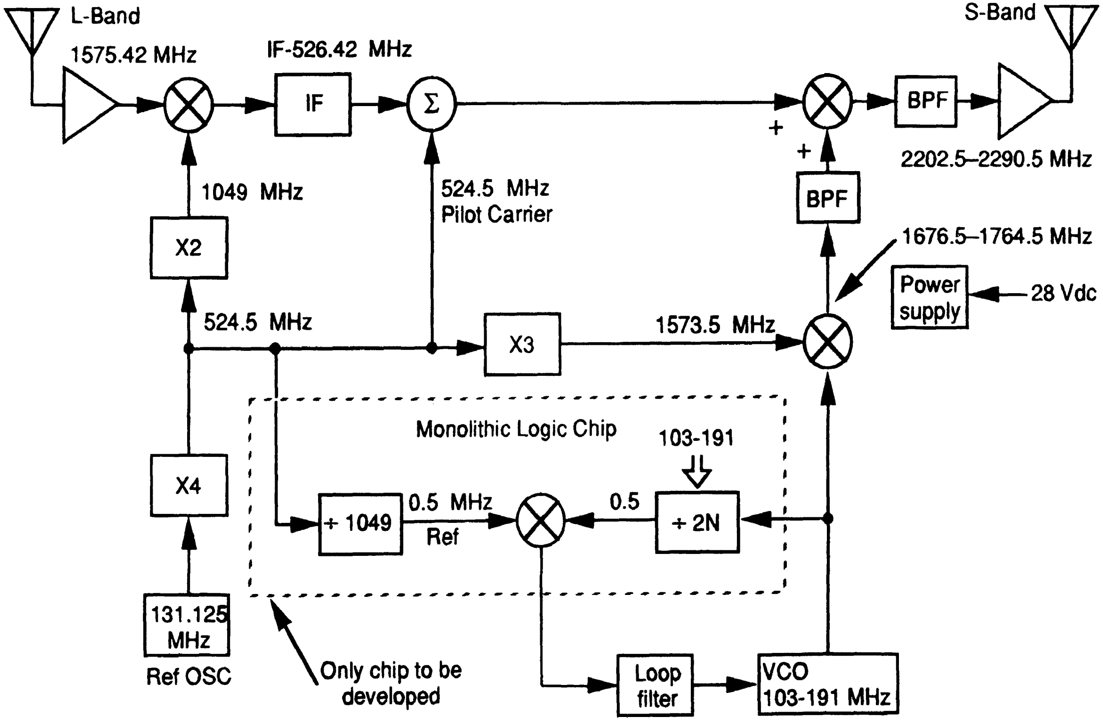


Fig. 7 RAJPO analog translator block diagram.

receives the C/A-code bandwidth portion of the rf spectrum, applies filtering and upconverts the captured spectrum to a fixed, (user-specified) S-band frequency; no signal detection or decoding is performed. The upconversion is coherently tied to a reference oscillator, also part of the translator. A pilot carrier, coherently related to the translator's reference oscillator, is added after the upconversion for TPS removal of the translator reference oscillator effects. The combined signal (S-band C/A code and pilot carrier) is then transmitted to the TPS.

2) *Current translators*: Two translator types have been built and demonstrated; i.e., the analog translator called the ballistic missile translator (BMT), and the dual frequency translator (DFT) developed for the Peacekeeper program. The dual frequency translator is a standard analog translator (L_1, C/A code) with an add-on module to translate a C/A-code bandwidth of the rf spectrum centered about the L_2 signal. Although the L_2 signal only contains the P code, it is possible to correlate a partially captured bandwidth, with only some loss in signal strength, by tracking the L_1 signal and processing the translated signals.

3) *Translator development*: Digital translators are similar to analog translators except that the translated signal is sampled at a high rate (4 million samples/s). This stream of sampled data is then digitized by an encryption device. However, the sampling losses are on the order of 5 dB over that of the analog translator development effort, and various studies have been done to investigate volume reduction. Current analog translators are about 40 in.^{3}. Proposals are being considered for 30, 20, and as small as 9 cubic inches.

Translator processing system—1) *Operation*: The TPS receives the translated (analog) GPS signals from one or more translators through range-owned telemetry antennas. The telemetry antennas must provide high gain to ensure a sufficient positive downlink signal-to-noise ratio (SNR) to preserve the translated GPS signal level. The translated GPS signals are input to a diversity receiver where the pilot carrier tone is tracked and removed, and the translated signals are downconverted to GPS L-band signals. The received translated signal is also predetect recorded. A reference receiver, part of the TPS, tracks all GPS satellite vehicles in view and provides data to the code-carrier tracker to aid in the translated signal processing. Figure 8 shows the block diagram of the translator processing system/vehicle tracker.

2) *Translator processing system development*: Substantial reductions in size and cost are planned for the next generation of TPS units. A multichannel TPS is planned to be housed in one of two racks. In addition, the analog tape recorders, no longer logistically supported by Ampex Corp., are planned to be replaced with solid state recording, using "flash" memory.

f. Ground transmitter (GT). A ground transmitter (GT) is a ground-based GPS satellite vehicle (SV) (sometimes called a "pseudolite") used on test and training ranges to augment/supplement the GPS constellation. Each GT transmits a GPS satellite-like signal that can be received by RAJPO-developed receivers. The actual transmitted signal consists of two carriers in phase quadrature centered at 1575.42 MHz (i.e., L_1), with each bi-phase modulated by the C/A code (one of 36 codes) and the P code (one of 37 codes). In this respect, the GT signal is identical to a satellite L_1 signal, although GTs do not broadcast L_2 (1227.6 MHz). The key difference between a GT and SV signal is the data content of the message

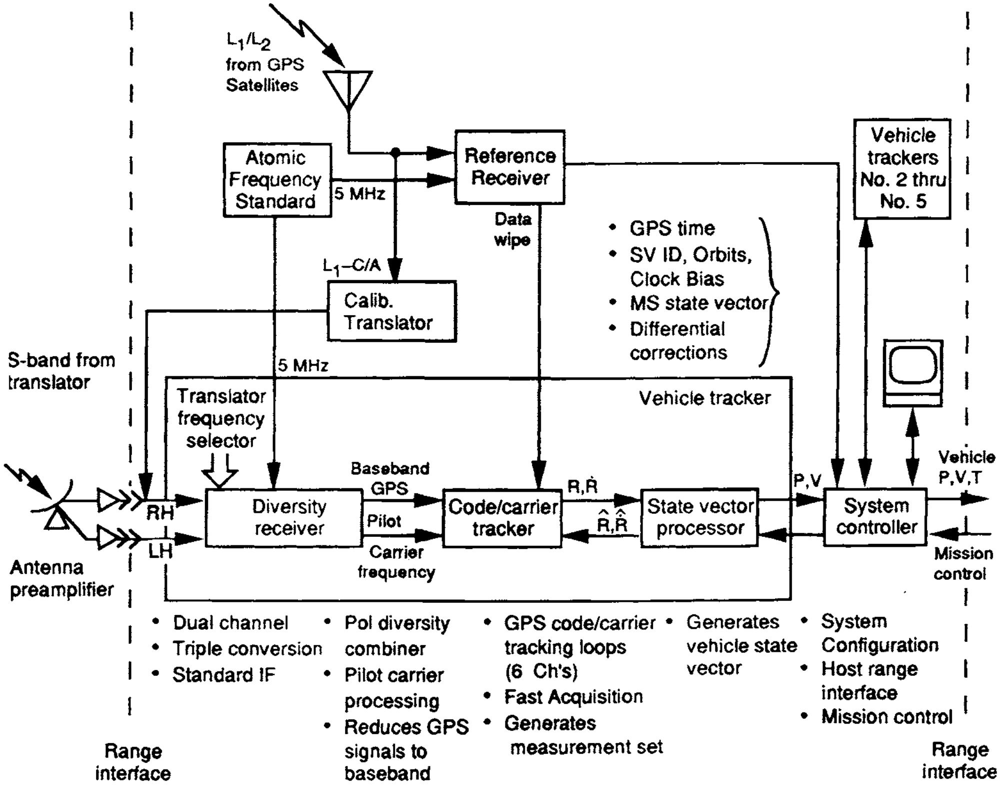


Fig. 8 Translator processing system/vehicle tracker block diagram.

datastream that modulo-two multiplies the codes. In particular, the ephemeris data must be different for a GT vs an SV.

RAJPO contracted Stanford Telecommunications, Inc. to develop GTs primarily for use with SDIO test programs. Ground transmitters are presently used in the Pacific to supplement the SV constellation. Their signals can be processed by the RAJPO receivers and the RAJPO TPS—the ground-based portion of the RAJPO GPS translator system. Figure 9 is an artist's rendition of a RAJPO model 5502 GT. Each GT simulates one GPS SV, and typically several GTs are employed over a test range to augment the GPS constellation or to account for line-of-sight blockage to an SV. GTs perform four basic functions: 1) synchronization to the GPS L_1 C/A- and P-code SV time signals—P code is primary; 2) synchronization to the GPS L_2 P-code time signal; 3) generation of a master timing reference compensated for first-order ionospheric time delay—using the L_1 and L_2 signals; and 4) simultaneous transmission of a simulated L_1 C/A- and P-code signal.

A typical GT scenario is provided in Fig. 10. The figure depicts the stand-alone operation of each GT for initial GPS system time synchronization with visible SVs and the subsequent transmission of GPS signals.

2. *Global Positioning Systems Joint Program Office*

The first differential GPS test was conducted by the JPO at the U.S. Army Yuma Proving Ground (YPG) in December 1979.[11] The Inverted Range Control Center (IRCC) was modified to operate as a differential GPS reference station, to compute the pseudorange corrections, and to transmit them to a test vehicle via the navigation message from a ground transmitter. The IRCC continued to operate as a reference station until 1987 and was used to monitor the space and control segments and to continue the development of differential GPS techniques.

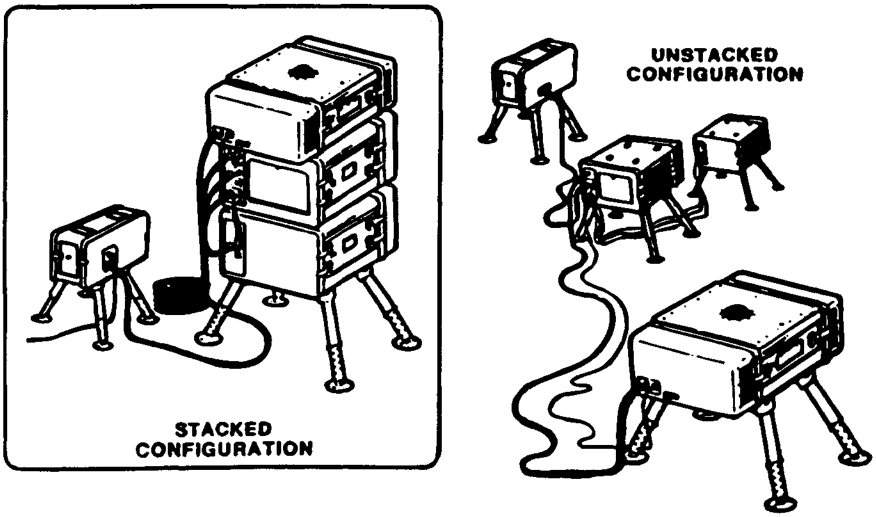


Fig. 9 RAJPO GPS ground transmitter—model 5502.

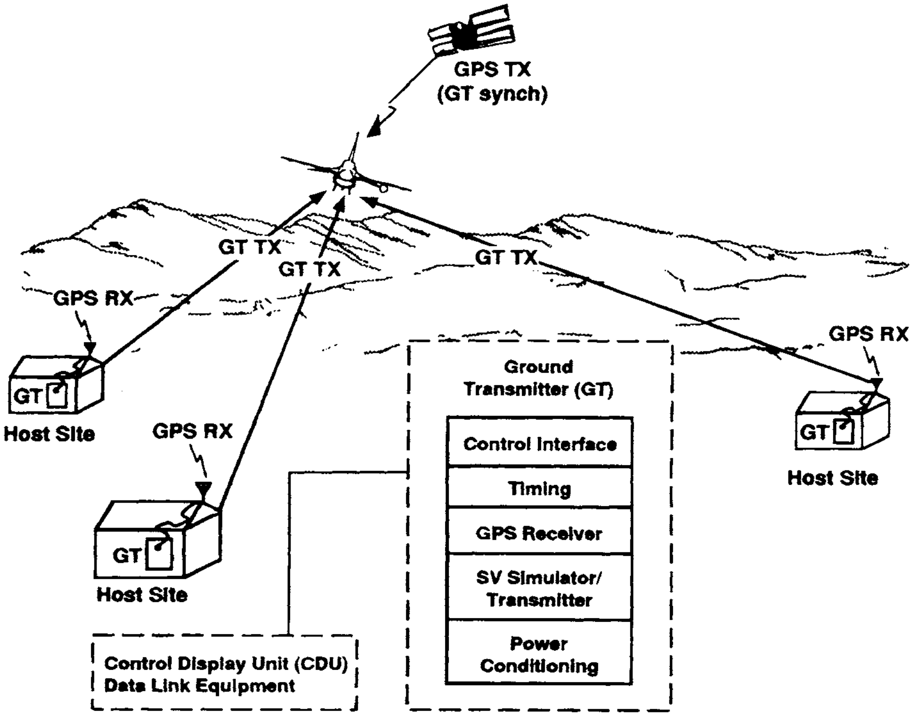


Fig. 10 Ground transmitter scenario.

During this same period, tests were conducted to validate the YPG range and to evaluate JPO Phase II GPS receivers as a reference system for range applications.

In 1985 a study was conducted to replace the IRCC with a dedicated GPS reference station. A system was built using TI-4100 GPS receivers and delivered to the JPO in 1987. An identical system was built and delivered to the B-2 Aircraft Combined Test Force (CTF) at Edwards Air Force Base and was used as the reference station to generate TSPI for test support in November 1987. A TI-4100 was used as the aircraft GPS receiver and TSPI was generated for flight test missions.

The TI-4100 reference receiver was replaced with a Collins 3A receiver in 1990. At that time, General Dynamics Services Company designed[12] and delivered one new reference station to the B-2 CTF and three systems to the JPO for test applications. Collins 3A receivers also replaced the TI-4100s in the flight test aircraft.

In 1991, responsibility for managing the four JPO reference stations called data analysis stations (DAS) was transferred to the 6585th Test Group's guidance test division, also known as the Central Inertial Guidance Test Division (CIGTF) at Holloman, Air Force Base, New Mexico. Three of the DGPS systems have been installed at Holloman Air Force Base, New Mexico, Edwards Air Force Base, California, and Melbourne, Florida to support DOD test programs. The fourth system is installed in a trailer and supports test programs on a mobile basis.

a. Equipment description.

Test platform—The GPS receivers used in the test vehicle are the Collins 3A or the Collins miniaturized airbone GPS receiver (MAGR). Raw measurement data from the RS-422 instrumentation port (IP) is recorded for postprocessing. Data are typically recorded on a PC buffer box (PCBB), which is either a 286/386 PC with a large hard disk or a digital tape recorder. If an analog tape recorder is available, the RS-422 digital data can be recorded on one channel and then downloaded after the mission to a PCBB. After each mission, the data are transported to the GPS reference station for processing and generation of TSPI data.

Ground station—The ground station configuration is shown in Fig. 11. It consists of four principle components: GPS receiver, GPS antenna, computer, and assorted input/output devices.

The GPS receiver is a Collins 3A, five-channel, two-frequency, P-code receiver modified to allow external control of tracking channels and for an external clock input. The raw pseudorange and delta range measurements and other required data from up to 12 satellites are transmitted to the computer for processing and recording. The antenna is a Dorne-Margolin.

The computer is a 80386-based system. Its real-time functions include control of the receiver, selection of satellites to be tracked, correction of measurements for propagation effects, and computation of the pseudorange corrections. In addition, the computer is used in a postprocessing mode to generate the final TSPI product.

The input/output devices are shown in Fig. 11. The primary data recording system is the Bernoulli removable 5 1/4 disk unit. The TSPI output for use by other agencies can be provided on either nine-track tape or Bernoulli disks. The printer is used to generate data products, plots, etc. for analysis.

B. Commercial Systems

The commercial industry has combined the use of differential GPS with low-cost C/A and P-code GPS receiver technology to develop small, lightweight, cost-effective, turn-key systems for range applications. The commercial vendors can provide either 1) complete turn-key systems that can be placed into operation immediately; or 2) hardware and software components that enable users to design a system to meet their requirements. Because the test and training applications have similar requirements, the training agencies are also taking advantage of the commercial equipment.

The generic GPS commercial range system block diagram is shown in Fig. 12. The GPS reference station tracks all visible satellites and pseudorange (PR) corrections are computed for each visible satellite and transmitted via the radio communications link to the mobile units. The mobile unit applies the appropriate PR corrections and performs a real-time computation to derive position and velocity. This solution is available for display in the test vehicle if required and is also transmitted back to the master control station for display and recording. The availability of the very accurate GPS time and the use of TDMA provides the capability to transmit data from 10–100 players (depending on the amount

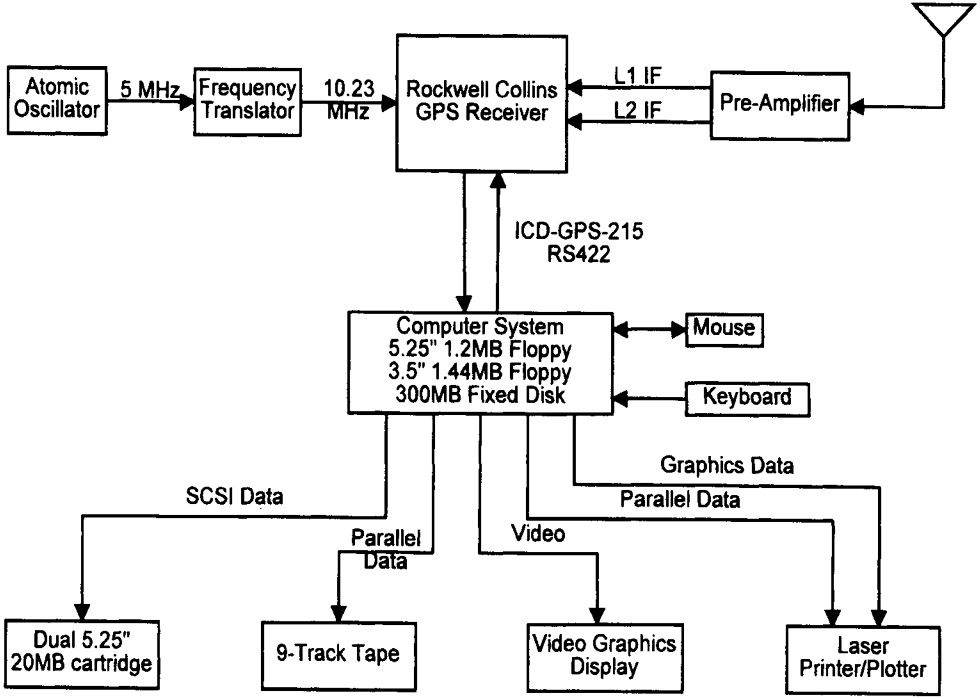


Fig. 11 Hardware configuration.

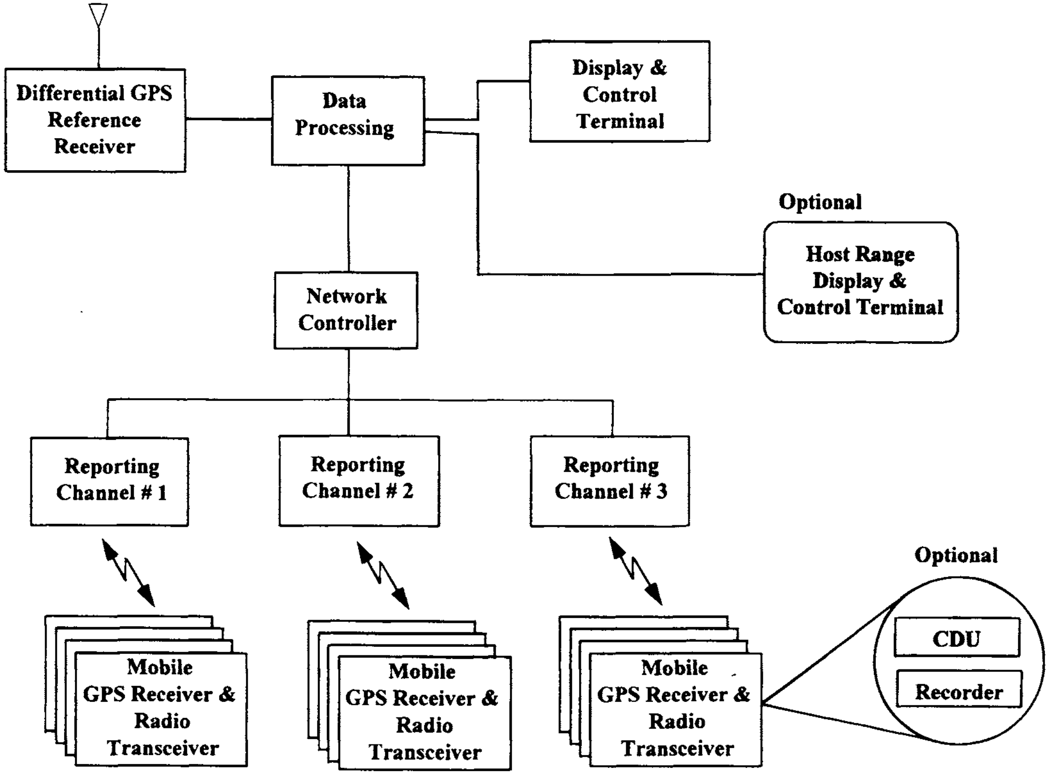

Differential GPS Reference Receiver
Data Processing
Display & Control Terminal
Optional
Host Range Display & Control Terminal
Network Controller
Reporting Channel # 1
Reporting Channel # 2
Reporting Channel # 3
Optional
CDU
Recorder
Mobile GPS Receiver & Radio Transceiver
Mobile GPS Receiver & Radio Transceiver
Mobile GPS Receiver & Radio Transceiver

Fig. 12 Generic commercial GPS range system.

of data) on one frequency. The use of multiple frequencies can increase the number of players by the number of channels available.

The U.S. Army YPG is operating a system with a capacity for 24 players to support positioning of aircraft, helicopters, and ground targets to evaluate airborne targeting sensors. The critical problem for the design of a DGPS system was to be able to collect data from ground targets in the rough desert terrain. Yuma Proving Ground is using a system developed by Trimble Navigation to support these requirements. The system uses a Trimble 4000RL differential reference station and six-channel C/A-code receivers for the mobile units. In order to meet the requirement to link data from ground vehicles in rough terrain, Trimble used off-the-shelf low-band vlf communications radios. GPS corrections are broadcast about every 10 s with mobile unit position reports scheduled or polled during the intervening period. In range operations, a base station collects player ID, position, and velocity of each participant and displays this information on a high-resolution color display on a digitized map background to support situation status in real time.

White Sands Missile Range has procured a 10-player GPS range system to support testing of a forward area Air-Defense Command, Control, and Intelligence System. The system was developed by SRI International using off-the-shelf GPS and communication radio equipment. SRI used a NavStar PLM/XR3 for the GPS reference station, Magnavox 4200 GPS receivers for the mobile GPS receivers, and Motorola VHF rf-Modems for the data link.

The positioning systems used for training applications are very similar. However, the total system is more complex because of the requirements for information on war gaming such as RTCA, probability of kill calculations, weapon system data, etc. A generic training system block diagram is shown in Fig. 13. Examples of systems currently deployed or being developed are briefly described, and references are provided.

Training systems currently in development, test, and deployment include the following. Simulated Area Weapons Effects–Radio Frequency (SAWE-RF) is a program that addresses indirect fire weapons, training of mounted and dismounted troops using computer simulated weapons, as well as the multiple integrated laser engagement system (MILES). The Phantom Run Instrumented MILES Enhancement (PRIME) system is being developed to enhance training for armored vehicles. The Army is working on a system that combines features of PRIME and SAWE-RF called Combat Maneuvering Training Center–Instrumentation System (CMTC–IS), primarily for armored vehicle training at the Hohenfels Training area in Germany. In addition, the Army is planning to develop a transportable system that combines all aspects of modern army warfare, including close air support and defense. This system, called Mobile Automated Instrumentation Suite (MAIS) will be designed to be deployed at any location worldwide and to be operational within 5 days. Magnavox MX 7100 and MX 4200 6-channel C/A-code receivers are used by most of these training systems as the differential GPS equipment. Details on these programs can be found in Refs. 13–15.

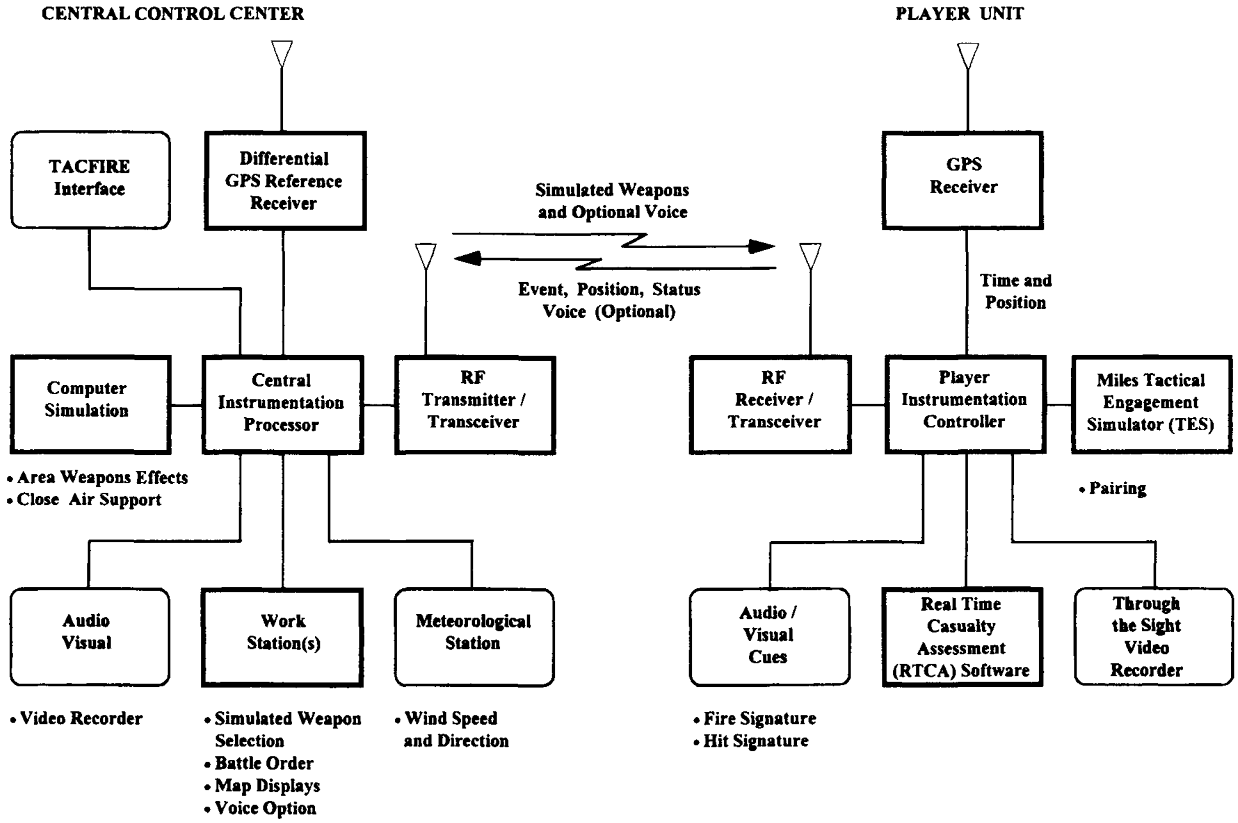

CENTRAL CONTROL CENTER
PLAYER UNIT
TACFIRE Interface
Differential GPS Reference Receiver
Simulated Weapons and Optional Voice
Event, Position, Status Voice (Optional)
GPS Receiver
Time and Position
Computer Simulation
Central Instrumentation Processor
RF Transmitter / Transceiver
RF Receiver / Transceiver
Player Instrumentation Controller
Miles Tactical Engagement Simulator (TES)
• Area Weapons Effects
• Close Air Support
• Pairing
Audio Visual
Work Station(s)
Meteorological Station
Audio / Visual Cues
Real Time Casualty Assessment (RTCA) Software
Through the Sight Video Recorder
• Video Recorder
• Simulated Weapon Selection
• Battle Order
• Map Displays
• Voice Option
• Wind Speed and Direction
• Fire Signature
• Hit Signature

Fig. 13 Generalized battlefield training instrumentation.

C. Data Links

The area that most limits the use of differential GPS in range applications is linking of corrections and/or TSPI to where it is required. The factors must be considered are the following: 1) data rate; 2) test vehicle dynamics; 3) size of area to be covered; 4) cost; 5) number of participants; and 6) DGPS method.

Data link requirements will be addressed by area size progressively. Diameters of areas considered will be 25–50 miles, 50–200 miles, 200–1000 miles, and greater than 1000 miles or what is termed wide-area differential. For the first case, the design issues are minimal, and as previously discussed, off-the-shelf communication radio equipment along with TDMA and use of multiple frequencies can accommodate hundreds of participants at a fairly reasonable cost.

For the 50–200 mile case, the rf line-of-sight limitations become a problem. The RAJPO is using ground relays to transmit data bidirectional from the ground station. The RAJPO data link system is a custom design to handle up to 200 players over these distances. The key word is custom, which results in a high-cost solution to the problem. For limited numbers of low dynamic players, a potential solution would be cellular telephones where coverage is available. Satellite communications is a solution described in the following paragraphs.

The cases of 200–1000 miles and over 1000 miles have the same data link problem but the potential need for additional reference stations comes into play. The most effective data link solution is satellite communications.[16–19] Cost, however, at this point is a limiting factor.

The radius of coverage for one differential station depends on several factors. A P-code reference station can provide coverage over a larger area than C/A-code systems because of the dual-frequency code-tracking capability. The 6585th Test Group at Holloman has verified differential GPS accuracies less than 5 m on test aircraft at distances of up to 600 miles. The coverage for C/A-code reference stations, however, is limited to approximately 50–100 miles. [A concept useful for large or nationwide test beds is wide area DGPS (WADGPS) as described in detail in Chapter 3 of this volume.]

A "network" concept for linking reference stations and generating differential corrections over relatively broad areas is also being studied. Pseudorange corrections (PRC) are measured at each reference station and then processed at a central location to generate corrections as a function of user location. The resultant is an "iso-PRC" contour map for each satellite. Because of the slowly changing error sources and change of the line-of sight vector to each satellite, the contour maps would have to be updated frequently.[20]

A series of tests to evaluate the use of a network of GPS reference stations as a source of differential corrections was conducted for the Burlington Northern Railroad. The tests were conducted over networks of 100, 200, and 300 miles. The results showed that the network concept can be used to cover large areas and achieve accuracy requirements required by the test range community.[21]

VI. Accuracy Performance

A. Position Accuracy

Test results from evaluation of C/A-code range systems against an accepted truth reference are very limited. The Joint Program Office has conducted limited

Table 2 Joint Program Office C/A code differential GPS test results

					Meters			
Date	No. Missions	Vehicle	Receiver	Code	VLEP	CEP	SEP	3 drms
8/87	5	Static	Ti 420	C/A	6.1	3.1	7.3	5.7
8/87	5	Truck	Ti 420	C/A	4.9	2.8	6.2	6.0
11/87	1	U-21	Tans 2Ch	C/A	3.2	4.7	6.5	6.2
1/89	4	Bac 1-11	Ti 420	C/A	3.7	5.0	7.5	8.1

VLEP, Vertical Linear Error Probable; CEP, Circular Error Probable; SEP, Spherical Error Probable; 3 drms, 3-dimensional root mean square.

testing of C/A-code receivers at YPG, and differential processing and analysis was performed on these receivers. A summary of those results are given in Table 2.

The JPO P-code differential GPS test support capability has undergone extensive testing at YPG under a variety of conditions. The YPG laser system was the truth reference for all tests. The results are summarized in Table 3.

The difference in the position accuracies between P code and C/A code seem to be approximately 3 m. P-code accuracies range from 2–4 m 3drms and C/A-code accuracies range from 6–8 m 3drms. If there is a conclusion to be made, it is that high-accuracy, high-dynamic test and training requirements shown in Table 1 will most likely require P-code systems, which requires more investment and complexity. On the other hand, the low-cost C/A-code systems can meet many of the land and low-dynamic requirements very cost effectively.

B. Velocity Accuracy

Validation of GPS velocity accuracies is even more of a problem because of the lack of accurate truth reference systems. The laser tracker velocity accuracy

Table 3 Joint Program Office P-code differential GPS test results

					Meters			
Date	No. Missions	Vehicle	Receiver	Code	VLEP	CEP	SEP	3 drms
11/85	1	Static	R/C 3A	P	1.3	1.3	2.1	2.5
10/86	4	Conv 440	TI INAV	P	1.1	2.3	2.8	2.7
11/88	1	B-52	R/C 3A	P	1.8	1.9	3.0	3.5
12/88	1	Truck	R/C 3A	P	1.3	1.1	1.9	2.4
1/89	4	Bac 1-11	R/C 3A	P	2.3	2.5	4.0	5.0
2/89	1	Rc-135	R/C 3A	P	1.8	2.0	3.1	3.6
5/89	1	Bus	R/C 3A	P	1.3	0.8	1.7	2.3
5/89	1	F-16	R/C 3A	P	1.4	1.1	2.0	3.0
6/89	1	Rc-135	R/C 3A	P	1.6	1.2	2.3	3.8
10/89	6	C-141	R/C 3A	P	1.1	1.5	2.1	3.1
9/90	8	T-39	R/C 3A	P	1.4	1.6	2.5	2.9

is only around 0.2–0.3 m/s, which is inadequate to evaluate the GPS specification accuracy of 0.1 m/s. Spot checks have been performed with specialized systems that have verified GPS velocity accuracies of <0.1 m/s.

In 1986, General Dynamics Services Company, the JPO support contractor at YPG was able to acquire Collins 3A data from an Army velocity accuracy test conducted by Draper Laboratories. The Aerial Profiling of Terrain System (APTS) was developed by Draper laboratory for the United States Geological Survey for unique mapping applications.[22] The APTS incorporates an inertial navigation system that produces position and velocity data. Laser ranging to surveyed retroreflectors on the ground provides a companion navigation system that removes the long-term increase of position errors attributable to drift, misalignment, and gravitational anomalies in the inertial solution, and ties that solution to a local geodetic coordinate system. Recorded data is postprocessed to yield very accurate position and velocity. One-sigma errors of postprocessed data are typically 1.0 cm and 0.3 mm/s during lock-on to a retroreflector; errors increase to 50 cm and 5 mm/s150 s after lock (assuming one retroreflector is acquired every 5 min).

During one of the aircraft missions, a Collins 3A receiver was aboard, and data were recorded. General Dynamics produced postmission differential GPS solutions for position and velocity using a reference station located at YPG. The differential solution was compared with the APTS truth trajectory processed by the Draper Laboratory. The standard deviation of the differences in three axis were the following:

East	0.03 m/s
North	0.03 m/s
Vertical	0.05 m/s

A second velocity verification test was conducted by the 6585th Test Group, Guidance Test Division, also know as the Central Inertial Guidance Test Facility at Holloman Air Force Base, New Mexico using the instrumented test track.[23] Low dynamic, constant velocity tests at from 20–35 mph were conducted using an unaided Collins 3A receiver. The standard deviation of the differences in three axis were the following:

East	0.03 m/s
North	0.03 m/s
Vertical	0.05 m/s

High velocity, 2–3*g* rocket tests were conducted also, but as expected, the unaided receiver velocity accuracies were much worse. Although the receiver maintained lock, the velocity errors were as much as 2 m/s at maximum acceleration.

VII. Future Developments

A. National Range

The concept of a national range using WADGPS concepts discussed earlier in this chapter has the potential for very cost effectively satisfying many DOD as well as civilian test and training requirements. The DOD SA/AS requirement placed on GPS complicates the implementation of an authorized P-code national range, but it could be done. The next step would be to determine how DOD and civilian applications could both use such a range and yet not violate any security aspects. The national range concept should be strongly considered as the next step in supporting test and training applications.

B. Kinematic Techniques

The other development required to meet the high-accuracy test and training requirements is carrier phase tracking technology. The civilian community is pursuing this area, and progress is good, but the host vehicle dynamics for most civilian applications are very low. Development is required to ensure continuous trajectories in an automated procedure for high-dynamic platforms.

References

[1] Blackwell, E. G., Cline, J. F., and Erb, E. A., "Technical and Economic Feasibility of Airborne and Satellite Instrumentation Systems to Augment National Test and Evaluation Resources," Final Rept., Contr. MDA903-78-C-0405, SRI International, Menlo Park, CA, July 1979.

[2] Reinhart, W. L., "Application of the Navstar Global Positioning System on Instrumented Ranges," Masters Thesis, Naval Postgraduate School, Monterey, CA, March 1981.

[3] Fredericksen, J. N., "Applicability of Navstar GPS to Test and Training," MITRE Corp. Rept., Contr. F19628-80-0001, May 1980.

[4] Sieg, W. D., "Applying GPS to Test Ranges," *ITEA Journal of Test and Evaluation,* Vol. 10, No. 2, 1989, p. 24.

[5] Wells, L. L., "Real-Time Missile Tracking with GPS," *Navigation,* Vol. 28, No. 3, 1984, p. 224.

[6] Mai, R. W., "Air and Ground Vehicle Tracking System," *Proceedings of the Fifth International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 16–18, 1992, p. 101.

[7] RTCM Paper 134-89/SC 104-68, "RTCM Recommended Standards for Differential Navstar GPS Service, Version 2.0," Jan. 1, 1990.

[8] Brown, A., Sward, W., Pickett, R., Greenberg, R., and Wildhagen, P., "Test Results of the Advanced Translator Processing System," *Proceedings of the Fourth International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 11–13, 1991, p. 573.

[9] McConnel, J. B., Greenberg, R. H., Pickett, R. B., Wildhagen, P. C., and Brown, A., "Advances in GPS Translator Technology," *Proceedings of the International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 27–29, 1989, p. 115.

[10] Sturza, M. A., and Brown, A. K., "Digital Translator Design Trades," *Proceedings of the Fifth International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 16–18, 1992, p. 687.

[11] Teasley, S. P., Hoover, W. M., and Johnson, C. R., "Differential GPS Navigation," Texas Instruments, Inc., Plans Symposium, Dec. 1980.

[12] Robbins, J. E., "Reference Trajectories From GPS Measurements," *Navigation,* Vol. 35, No. 1, 1988, p. 89.

[13]Eastwood, R. A., and Sharpe, R. T. "The Use of GPS to Enhance the Military Training Environment," Paper presented at the NATO GPS Symposium, Brussels, Belgium, Nov. 20–22, 1990.

[14]Peters, R. L., and Lewis, K. M., "Use of GPS as the Position Location Subsystem for the Army's Prime Training Range System," *Proceedings of the Fourth International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 11–13, 1991, p. 593.

[15]Truog, B., and Ravenis, J., "Combat Manuever Training Center (CMTC) High Dynamic Player Integration," *Proceedings of the Fourth International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 11–13, 1991, p. 649.

[16]Blanchard, W. F., "Differential GPS Using a Dedicated INMARSAT Satellite Data Link," *Proceedings of the Third International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 19–21, 1990, p. 237.

[17]Nagle, J. R., "Wide Area Differential Corrections (WADC) from Global Beam Satellites," *Proceedings of the IEEE 0-7803-0468-3/92,* Institute of Electrical and Electronics Engineers, New York, 1992, p. 383.

[18]Slack, E. R., "Towards a Global Differential Service," *Proceedings of the Third International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 19–21, 1990, p. 323.

[19]Zachmann, G. W., "Differential GPS Transmissions By Geostationary L-Band Satellites," *Sea Technology,* Vol. 31, No. 5, 1990, p. 57.

[20]Loomis, P., Sheynblatt, L., and Mueller, T., "Differential GPS Network Design," *Proceedings of the Fourth International Technical Meeting of the Satellite Division of the ION,* Institute of Navigation, Washington, DC, Sept. 11–13, 1991, p. 511.

[21]Robbins, J., "Evaluation of Differential GPS Capabilities for Railroad Positioning Requirements," Paper presented at the DGPS 91 Symposium, Braunschweig, Germany, 1991.

[22]Greenspan, R. L., "APTS/GPS Measurement Task," Draper Laboratory Final TR, Cambridge, MA, Feb. 1986.

[23]MSD-DP-90-04, "Navstar GPS User Equipment Sled Test," Munitions Systems Division, Data Package, March 1990.

AUTHOR INDEX

SUBJECT INDEX